

IBM SPSS Regression 19



Note: Before using this information and the product it supports, read the general information under Notices a pag. 46.

This document contains proprietary information of SPSS Inc, an IBM Company. It is provided under a license agreement and is protected by copyright law. The information contained in this publication does not include any product warranties, and any statements provided in this manual should not be interpreted as such.

When you send information to IBM or SPSS, you grant IBM and SPSS a nonexclusive right to use or distribute the information in any way it believes appropriate without incurring any obligation to you.

© **Copyright SPSS Inc. 1989, 2010.**

Prefazione

IBM® SPSS® Statistics è un sistema completo per l'analisi dei dati. Il modulo aggiuntivo opzionale Regression include le tecniche di analisi aggiuntive descritte nel presente manuale. Il modulo aggiuntivo Regression deve essere usato con il modulo Core SPSS Statistics in cui è completamente integrato.

Informazioni su SPSS Inc., una società del gruppo IBM

SPSS Inc., una società del gruppo IBM, è fornitore leader mondiale nel settore del software e delle soluzioni per l'analisi predittiva. L'offerta completa dei prodotti dell'azienda (raccolta di dati, statistica, modellazione e distribuzione) consente di acquisire i comportamenti e le opinioni delle persone, prevedere i risultati delle future interazioni con i clienti ed elaborare questi dati integrando le analitiche nelle procedure aziendali. Le soluzioni SPSS Inc. consentono la gestione di attività interconnesse all'interno dell'intera organizzazione, con particolare attenzione alla convergenza di analitiche, architettura IT e procedure aziendali. Clienti commerciali, istituzionali e accademici di tutto il mondo si affidano alla tecnologia SPSS Inc. ottenendo un vantaggio competitivo in termini di attrazione, mantenimento e ampliamento della base clienti, riducendo al contempo frodi e rischi. SPSS Inc. è stata acquisita da IBM nell'ottobre 2009. Per ulteriori informazioni, visitare il sito <http://www.spss.com>.

Supporto tecnico

Ai clienti che richiedono la manutenzione, viene messo a disposizione un servizio di supporto tecnico. I clienti possono contattare il supporto tecnico per richiedere assistenza per l'utilizzo dei prodotti SPSS Inc. o per l'installazione di uno degli ambienti hardware supportati. Per il supporto tecnico, visitare il sito Web di SPSS Inc. all'indirizzo <http://support.spss.com> o contattare la filiale del proprio paese indicata nel sito Web all'indirizzo <http://support.spss.com/default.asp?refpage=contactus.asp>. Ricordare che durante la richiesta di assistenza sarà necessario fornire i dati di identificazione personali, i dati relativi alla propria società e il numero del contratto di manutenzione.

Servizio clienti

Per informazioni sulla spedizione o sul proprio account, contattare la filiale nel proprio paese, indicata nel sito Web all'indirizzo <http://www.spss.com/worldwide>. Tenere presente che sarà necessario fornire il numero di serie.

Corsi di formazione

SPSS Inc. organizza corsi di formazione pubblici e onsite che includono esercitazioni pratiche. Tali corsi si terranno periodicamente nelle principali città. Per ulteriori informazioni sui corsi, contattare la filiale nel proprio paese, indicata nel sito Web all'indirizzo <http://www.spss.com/worldwide>.

Pubblicazioni aggiuntive

I documenti *SPSS Statistics: Guide to Data Analysis*, *SPSS Statistics: Statistical Procedures Companion* e *SPSS Statistics: Advanced Statistical Procedures Companion*, scritti da Marija Norušis e pubblicati da Prentice Hall sono disponibili come materiale supplementare consigliato. Queste pubblicazioni descrivono le procedure statistiche nei moduli SPSS Statistics Base, Advanced Statistics e Regression. Utili sia come guida iniziale all'analisi dei dati che per applicazioni avanzate, questi manuali consentono di ottimizzare l'utilizzo delle funzionalità presenti nell'offerta IBM® SPSS® Statistics. Per ulteriori informazioni, inclusi contenuti delle pubblicazioni e capitoli di esempio, visitare il sito Web dell'autrice: <http://www.norusis.com>

Contenuto

1 Scelta di una procedura per la regressione logistica binaria 1

2 Regressione logistica 3

Regressione logistica: Imposta valore	5
Metodi di selezione delle variabili della regressione logistica	5
Regressione logistica: Definisci variabili categoriche	6
Regressione logistica: Salva nuove variabili	7
Regressione logistica: Opzioni	9
Opzioni aggiuntive del comando LOGISTIC REGRESSION	10

3 Regressione logistica multinomiale 11

Regressione logistica multinomiale	13
Costruisci termini	14
Regressione logistica multinomiale: Categoria riferimento	15
Regressione logistica multinomiale: Statistiche	16
Regressione logistica multinomiale: Criteri di convergenza	17
Regressione logistica multinomiale: Criteri di convergenza	18
Regressione logistica multinomiale: Salva	20
Opzioni aggiuntive del comando NOMREG	20

4 Analisi Probit 21

Analisi probit: Definisci intervallo	23
Analisi probit: Opzioni	23
Opzioni aggiuntive del comando PROBIT	24

5 Regressione non lineare 25

Logica condizionale (Regressione non lineare)	26
---	----

Regressione non lineare: Parametri	27
Modelli comuni della Regressione non lineare	28
Regressione non lineare: Funzione di perdita	29
Regressione non lineare: Vincoli sui parametri	30
Regressione non lineare: Salva nuove variabili	31
Regressione non lineare: Opzioni	31
Interpretazione dei risultati della regressione non lineare	32
Opzioni aggiuntive del comando NLR	33
6 <i>Stima del peso</i>	34
Minimi quadrati ponderati (WLS): Opzioni	36
Opzioni aggiuntive del comando WLS	36
7 <i>Minimi quadrati a 2 stadi</i>	37
Minimi quadrati a 2 stadi: Opzioni	39
Opzioni aggiuntive del comando 2SLS	39
<i>Appendici</i>	
A <i>Schemi di codifica delle variabili categoriali</i>	40
Deviazione	40
Semplice	41
Helmert	42
Differenza	42
Polinomiale	43
Ripetuto	43
Contrasto speciale	44
Indicatore	45

B Notices

46

Indice

48

Scelta di una procedura per la regressione logistica binaria

I modelli di regressione logistica binaria possono essere adattati mediante la procedura Regressione logistica oppure la procedura Regressione logistica multinomiale. Entrambe includono opzioni non disponibili nell'altra. Un'importante differenza a livello teorico è rappresentata dal fatto che la procedura di regressione logistica fornisce valori stimati, residui, statistiche di influenza e test sulla bontà dell'adattamento basati sui dati del singolo caso, indipendentemente dalla modalità di inserimento dei dati e dal fatto che il numero di modelli covariati sia o meno inferiore a quello totale dei casi. La procedura di regressione logistica multinomiale aggrega invece i casi internamente, in modo da formare sottopopolazioni con modelli covariati identici per gli stimatori e fornisce quindi valori stimati, residui e test sulla bontà dell'adattamento basati su tali sottopopolazioni. Se tutti gli stimatori sono di tipo categoriale o se a tutti gli stimatori continui è possibile passare solo un numero limitato di valori — per cui a ogni modello covariato diverso sono associati più casi separati — la procedura basata sulle sottopopolazioni è in grado di produrre test sulla bontà.

Regressione logistica rende disponibili le seguenti funzioni univoche:

- Test di bontà dell'adattamento di Hosmer-Lemeshow per il modello
- Analisi per passi
- Contrasti per definire la parametrizzazione del modello
- Punti di divisione alternativi per la classificazione
- Grafici di classificazione
- Modello adattato a una serie di casi per produrre una serie di casi
- Salvataggio di valori stimati, residui e statistiche di influenza

Regressione logistica multinomiale rende disponibili le seguenti funzionalità univoche:

- Test chi-quadrato di Pearson e della devianza sulla bontà dell'adattamento del modello
- Definizione di sottopopolazioni per raggruppare i dati per i test sulla bontà dell'adattamento
- Elenco di frequenze, frequenze attese e residui per sottopopolazione
- Correzione delle stime della varianza per la sovradisersione
- Matrice di covarianza delle stime dei parametri
- Test delle combinazioni lineari dei parametri

- Esplicitazione dei modelli nidificati
- Adattamento di corrispondenze 1 a 1 di modelli di regressione logistica condizionale mediante variabili differenziate

Regressione logistica

La regressione logistica risulta utile quando si desidera prevedere la presenza o l'assenza di una caratteristica o di un risultato in base ai valori di un insieme di variabili stimatore. È simile al modello di regressione lineare ma si adatta ai modelli in cui la variabile dipendente è dicotomica. È possibile utilizzare i coefficienti di regressione logistica per stimare i rapporti odd per ogni variabile indipendente nel modello. La regressione logistica è applicabile a un numero più elevato di ricerche rispetto all'analisi discriminante.

Esempio. Quali caratteristiche dello stile di vita rappresentano fattori di rischio per le malattie cardiovascolari (CHD)? In un campione di pazienti in cui è stato verificato lo stato di fumatore, la dieta, l'esercizio, l'uso di alcolici e lo stato delle malattie cardiovascolari (CHD), è possibile costruire un modello utilizzando le quattro variabili relative allo stile di vita per prevedere la presenza o l'assenza di CHD in un campione di pazienti. È quindi possibile utilizzare il modello per ottenere le stime dei rapporti odd per ogni fattore e per sapere, ad esempio, quante probabilità in più hanno i fumatori di sviluppare malattie CHD rispetto ai non fumatori.

Statistiche. Per ogni analisi: casi totali, casi selezionati, casi validi. Per ogni variabile categoriale: codifica del parametro. Per ogni passo: variabili inserite o rimosse, cronologia delle iterazioni, $-2 \log$ verosimiglianza, bontà di adattamento, test di bontà dell'adattamento di Hosmer-Lemeshow, modello chi-quadrato, chi-quadrato del miglioramento, tabella di classificazione, correlazioni tra variabili, grafico dei gruppi osservati e delle probabilità previste e chi-quadrato residuo. Per ogni variabile nell'equazione: coefficiente (B), errore standard di B , statistica di Wald, rapporto odd stimato ($\exp(B)$), intervallo di confidenza per $\exp(B)$, verosimiglianza se il termine è rimosso dal modello. Per ogni variabile non presente nell'equazione: statistica di punteggio. Per ogni caso: gruppo osservato, probabilità stimate, gruppo stimato, residuo, residuo standardizzato.

Metodi. È possibile stimare i modelli utilizzando l'inserimento del blocco di variabili o uno dei seguenti metodi stepwise: avanti condizionale, avanti LR, avanti Wald, indietro condizionale, indietro LR o indietro Wald.

Dati. La variabile dipendente deve essere dicotomica. Le variabili indipendenti possono essere a livello di intervallo o categoriali. Se sono categoriali, devono essere codificate come fittizie o indicatori (esiste un'opzione nella procedura per ricodificare le variabili categoriali in modo automatico).

Assunzioni. La regressione logistica non si basa sulle ipotesi di distribuzione allo stesso modo dell'analisi discriminante. È tuttavia possibile che la soluzione ottenuta sia più stabile se le variabili stimatore di previsione hanno una distribuzione normale multivariata. Inoltre, come con altre forme di regressione, la multicollinearità tra le variabili stimatore può portare a stime distorte e ad inflazione di errori standard. La procedura risulta più efficace quando il gruppo di appartenenza è una vera variabile categoriale. Se il gruppo di appartenenza è basato sui valori di una variabile continua (ad esempio, "QI alto" in contrapposizione a "QI basso"), è necessario

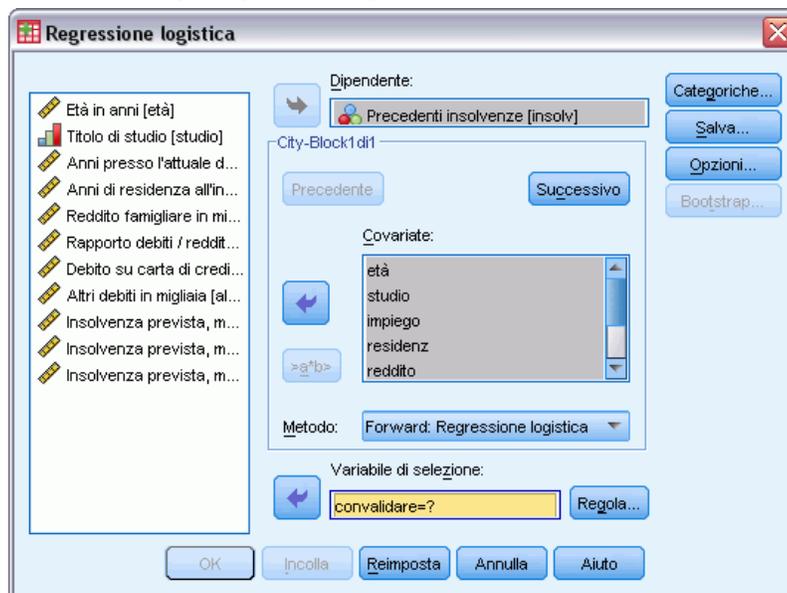
prendere in considerazione l'utilizzo della regressione lineare per ottenere informazioni più complete dalla variabile continua.

Procedure correlate. Utilizzare la procedura Grafico a dispersione per visualizzare i dati per la multicollinearità. Se le ipotesi della matrice di normalità multivariata e di varianza-covarianza hanno esito positivo, è possibile ottenere una soluzione più rapida utilizzando la procedura Analisi discriminante. Se tutte le variabili stimatore sono categoriali, è inoltre possibile utilizzare la procedura Loglineare. Se la variabile dipendente è continua, utilizzare la procedura Regressione lineare. È possibile utilizzare la procedura della curva ROC per rappresentare le probabilità salvate con la procedura di Regressione logistica.

Per ottenere un'analisi Regressione logistica

- Dai menu, scegliere:
Analizza > Regression > Logistica binaria...

Figura 2-1
Finestra di dialogo Regressione logistica



- Selezionare una variabile dipendente dicotomica. La variabile può essere di tipo numerico o stringa.
- Selezionare una o più covariate. Per includere i termini di interazione, selezionare tutte le variabili coinvolte nell'interazione e quindi selezionare >a*b>.

Per inserire le variabili nei gruppi (**blocchi**), selezionare le covariate per un blocco e quindi fare clic su Successivo per specificare un nuovo blocco. Ripetere la procedura fino a quando sono stati specificati tutti i gruppi.

In alternativa, è possibile selezionare i casi per l'analisi. Scegliere una variabile di selezione e fare clic su Imposta valore.

Regressione logistica: Imposta valore

Figura 2-2
Finestra di dialogo Regressione logistica: Imposta valore



Nella stima del modello verranno inseriti i casi definiti dalla regola di selezione impostata. Se, ad esempio, è stata selezionata una variabile uguale a e specificato il valore 5, solo i casi per cui la variabile selezionata ha un valore uguale a 5 sono inclusi nella stima del modello.

Le statistiche e i risultati della classificazione vengono generati sia per i casi selezionati che per i casi non selezionati. In questo modo viene messo a disposizione un meccanismo per la classificazione dei nuovi casi basato sui dati preesistenti o per il partizionamento dei dati in sottoinsiemi di esempio e prova utilizzabile per eseguire la convalida del modello generato.

Metodi di selezione delle variabili della regressione logistica

La selezione del metodo consente di specificare come vengono inserite nell'analisi le variabili indipendenti. Utilizzando diversi metodi, è possibile creare molteplici modelli di regressione dallo stesso insieme di variabili.

- **Invio.** Una procedura per la selezione delle variabili nella quale tutte le variabili di un blocco sono inserite in un unico passo.
- **Avanti: Condizionale.** Metodo di selezione per passi con test di inserimento basato sulla significatività della statistica di punteggio e con test di rimozione basato sulla probabilità di un rapporto di verosimiglianza fissato in base a stime condizionali dei parametri.
- **Avanti: Rapporto di verosimiglianza.** Selezione per passi con test di inserimento basato sulla significatività della statistica di punteggio e con test di rimozione basato sulla probabilità di un rapporto di verosimiglianza fissato in base a stime di massima verosimiglianza parziali.
- **Avanti: Wald.** Selezione per passi con test di inserimento basato sulla significatività della statistica di punteggio e con test di rimozione basato sulla probabilità della statistica di Wald.
- **Eliminazione all'indietro (Condizionale).** Selezione per passi all'indietro. Il test di rimozione è basato sulla probabilità del rapporto di verosimiglianza basato sulle stime condizionali dei parametri.
- **Eliminazione all'indietro: (Rapporto di verosimiglianza).** Selezione per passi all'indietro. Il test di rimozione è basato sulla probabilità del rapporto di verosimiglianza basato sulle stime della massima verosimiglianza parziale.
- **Eliminazione all'indietro (Wald).** Selezione per passi all'indietro. Il test di rimozione è basato sulla probabilità della statistica di Wald.

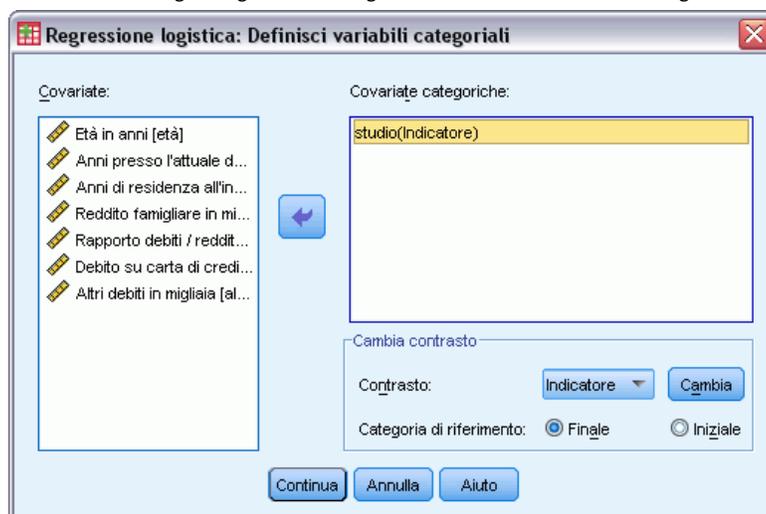
I valori di significatività dell'output si basano sull'adattamento di un singolo modello. Pertanto, i valori di significatività in genere non sono validi quando viene utilizzato un metodo stepwise.

Tutte le variabili indipendenti selezionate vengono aggiunte a un solo modello di regressione. È tuttavia possibile specificare diversi metodi di inserimento per diversi sottoinsiemi di variabili. Ad esempio, è possibile inserire un blocco di variabili nel modello di regressione utilizzando la selezione per passi e un secondo blocco utilizzando la selezione in avanti. Per aggiungere un secondo blocco di variabili a un modello di regressione, fare clic su Avanti.

Regressione logistica: Definisci variabili categoriche

Figura 2-3

Finestra di dialogo Regressione logistica: Definisci variabili categoriche



È possibile specificare i dettagli relativi alla modalità di gestione delle variabili categoriali della procedura di regressione logistica:

Covariate. Contiene un elenco di tutte le covariate specificate in qualsiasi strato della finestra di dialogo principale, da sole o come parte di un'interazione. Se alcune di queste variabili sono stringa o categoriali, è possibile utilizzarle solo come covariate categoriche.

Covariate categoriche. Elenca le variabili identificate come categoriali. Ogni variabile include una notazione tra parentesi che indica la codificazione di contrasto da utilizzare. Variabili stringa (codificate con il simbolo < dopo i relativi nomi) sono già presenti nell'elenco Covariate categoriche. Selezionare altre covariate categoriche dall'elenco di covariate e spostarle nell'elenco di covariate categoriche.

Cambia. Consente di modificare il metodo di contrasto. I metodi di contrasto disponibili sono:

- **Indicatore.** I contrasti indicano l'appartenenza o la non appartenenza alla categoria. La categoria di riferimento è rappresentata nella matrice di contrasto come una riga di zero.
- **Semplice.** Tutte le categorie della variabile stimatore (eccetto la categoria di riferimento) vengono confrontate con la categoria di riferimento.

- **Differenza.** Tutte le categorie della variabile stimatore (eccetto la prima categoria) vengono confrontate con l'effetto medio delle categorie precedenti. Sono noti anche come contrasti inversi di Helmert.
- **Helmert.** Tutte le categorie della variabile stimatore (eccetto l'ultima categoria) vengono confrontate con l'effetto medio delle categorie successive.
- **Ripetuto.** Tutte le categorie della variabile stimatore (eccetto la prima categoria) vengono confrontate con la categoria che le precede.
- **Polinomiale.** Contrasti polinomiali ortogonali. Si presume che le categorie siano equamente distanziate. I contrasti polinomiali sono disponibili solo per le variabili numeriche.
- **standard.** Tutte le categorie della variabile stimatore (eccetto la categoria di riferimento) vengono confrontate con l'effetto globale.

Se si seleziona Deviazione, Semplice o Indicatore, selezionare Iniziale o Finale come categoria di riferimento. Si noti che il metodo non viene effettivamente modificato finché non si fa clic su Cambia.

Le covariate di tipo stringa devono essere covariate categoriche. Per rimuovere una variabile di tipo stringa dall'elenco Covariate categoriche, è necessario rimuovere tutti i termini contenenti la variabile dall'elenco di covariate nella finestra di dialogo principale.

Regressione logistica: Salva nuove variabili

Figura 2-4

Finestra di dialogo Regressione logistica: Salva nuove variabili



È possibile salvare i risultati della regressione logistica come nuove variabili nel file di dati attivo

Valori attesi. Consente di salvare i valori previsti dal modello. Le opzioni disponibili sono Probabilità e Gruppo di appartenenza.

- **Probabilità.** Salva per ogni caso la probabilità stimata di occorrenza dell'evento. Una tabella nell'output visualizza il nome e il contenuto delle nuove variabili.
- **Gruppo di appartenenza previsto.** Il gruppo con la massima probabilità a posteriori, in base a punteggi discriminanti. Indica il gruppo a cui il modello assegnerebbe ciascun caso.

Influenza. Consente di salvare i valori dalle statistiche che misurano l'influenza dei casi sui valori attesi. Le opzioni disponibili sono di Cook, Valori di influenza e DiffBeta.

- **Di Cook (Regressione logistica: salva).** Il corrispondente della regressione logistica della statistica di Cook. Una misura di quanto cambierebbero i residui di tutti i casi se un particolare caso fosse escluso dal calcolo dei coefficienti di regressione.
- **Valore d'influenza.** L'influenza relativa di ogni osservazione sull'adattamento del modello.
- **Differenza in beta.** Variazione del coefficiente di regressione quando un caso particolare viene eliminato dall'analisi. Viene calcolato un valore per ogni termine del modello, incluso il termine costante.

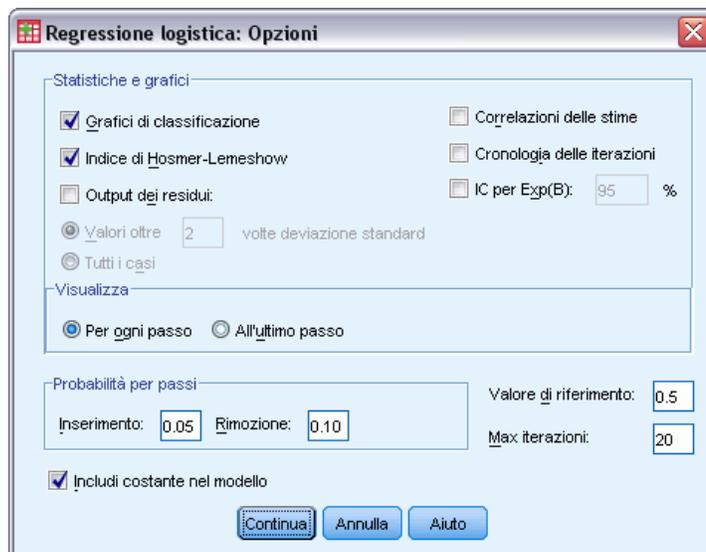
Residui. Consente di salvare i residui. Le opzioni disponibili sono Non standardizzati, Logit, Studentizzati, Standardizzati e Devianza.

- **Residui non standardizzati.** La differenza tra un valore osservato e il valore stimato dal modello.
- **Residuo logit.** Il residuo per il caso se viene stimato nella scala logit. Il residuo logit è il residuo diviso per la probabilità attesa e moltiplicato per 1 meno la probabilità attesa.
- **Residuo studentizzato (Regressione Logistica).** La variazione della devianza nel modello se un caso viene escluso.
- **Residui standardizzati.** Il residuo diviso per una stima della deviazione standard. Il residuo standardizzato, conosciuto anche come residuo di Pearson, ha media 0 e deviazione standard 1.
- **Devianza.** Residui basati sulla devianza del modello.

Esporta informazioni modello in file XML. Le stime dei parametri e, se si desidera, le relative covarianze vengono esportati nel file specificato in formato XML (PMML). È possibile utilizzare questo file di modello per applicare le informazioni del modello ad altri file di dati per il calcolo del punteggio.

Regressione logistica: Opzioni

Figura 2-5
Finestra di dialogo Regressione logistica: Opzioni



È possibile specificare le opzioni per l'analisi della regressione logistica:

Statistiche e grafici. Consente di richiedere statistiche e grafici. Le opzioni disponibili sono Grafici di classificazione, Indice di Hosmer-Lemeshow, Output dei residui, Correlazioni delle stime, Cronologia iterazioni e IC per $\exp(B)$. Selezionare una delle alternative nel gruppo Visualizza per visualizzare le statistiche e i grafici per ogni iterazione oppure solo per il modello finale.

- **Test di bontà dell'adattamento di Hosmer-Lemeshow.** Questa statistica è più robusta dei test di bontà dell'adattamento tipicamente usati nella regressione logistica, specialmente per i modelli con covariate continue e per piccoli campioni. Raggruppa i casi in decili di rischio e confronta la probabilità osservata di ogni decile con la corrispondente probabilità attesa

Probabilità per Stepwise. Consente di controllare i criteri in base ai quali le variabili sono inserite e rimosse dall'equazione. È possibile specificare i criteri per l'inserimento o la rimozione di variabili.

- **Probabilità per passi.** Nel modello viene inserita una variabile se la probabilità della sua statistica di punteggio è inferiore al valore di inserimento, mentre viene rimossa se la probabilità è superiore al valore di rimozione. Specificare valori di inserimento e rimozione positivi per annullare le impostazioni di default. Il valore di inserimento deve essere inferiore al valore di rimozione.

Valore di riferimento. Consente di determinare il punto di divisione per i casi da classificare. I casi con valori attesi che superano il valore di riferimento sono classificati come positivi, mentre i casi con valori previsti minori del valore di riferimento sono classificati come negativi. Per modificare il valore predefinito, inserire un valore compreso tra 0,01 e 0,99.

Massimo numero di iterazioni. Consente di modificare il numero massimo di volte in cui il modello procede all'iterazione prima di chiudere.

Includi costante nel modello. Consente di indicare se il modello può includere un termine costante. Se disattivata, il termine costante sarà uguale a 0.

Opzioni aggiuntive del comando LOGISTIC REGRESSION

Il linguaggio della sintassi dei comandi consente inoltre di:

- Identificare l'output per casi in base ai valori o alle etichette di variabile di una variabile.
- Controllare la spaziatura fra i rapporti di iterazione. Anziché stampare le stime dei parametri dopo ogni iterazione, è possibile richiederle dopo ciascuna n -esima iterazione.
- Modificare i criteri per terminare l'iterazione e per controllare la ridondanza.
- Specificare una lista di variabili per gli elenchi per casi.
- Risparmiare memoria conservando i dati relativi a ogni gruppo di file suddiviso in base alla variabile categoriale in un file scratch esterno durante l'elaborazione.

Vedere *Command Syntax Reference* per informazioni dettagliate sulla sintassi.

Regressione logistica multinomiale

La regressione logistica multinomiale è utile per le situazioni in cui si desidera classificare i soggetti in base ai valori di un insieme di variabili indipendenti. Si tratta di un tipo di regressione simile alla regressione logistica, ma è più generale in quanto la variabile dipendente non è limitata a due categorie.

Esempio. Uno studio cinematografico, per lanciare i suoi film in modo più efficace, vuole prevedere quali sono i tipi di film preferiti dai frequentatori di cinema. Utilizzando una regressione logistica multinomiale, lo studio può determinare l'influenza di età, sesso e stato civile sul tipo di film che un individuo preferisce. Lo studio potrà quindi indirizzare la campagna pubblicitaria di un determinato film verso un gruppo di spettatori più probabili.

Statistiche. Cronologia iterazioni, coefficienti parametrici, matrici di covarianza e correlazione asintotica, test del rapporto di verosimiglianza per gli effetti del modello e parziali, $-2 \log$ verosimiglianza. Chi-quadrato di Pearson e della devianza per la bontà dell'adattamento. R^2 di Cox e Snell, di Nagelkerke e di McFadden. Classificazione: confronto tra frequenze osservate e attese per categoria di risposta. Tavole di contingenza: frequenze osservate e attese (con residui) e proporzioni per modello covariata e categoria di risposta.

Metodi. Un modello logit multinomiale è adatto al modello fattoriale completo o a un modello specificato dall'utente. La stima dei parametri è eseguita attraverso un algoritmo iterativo di massima verosimiglianza.

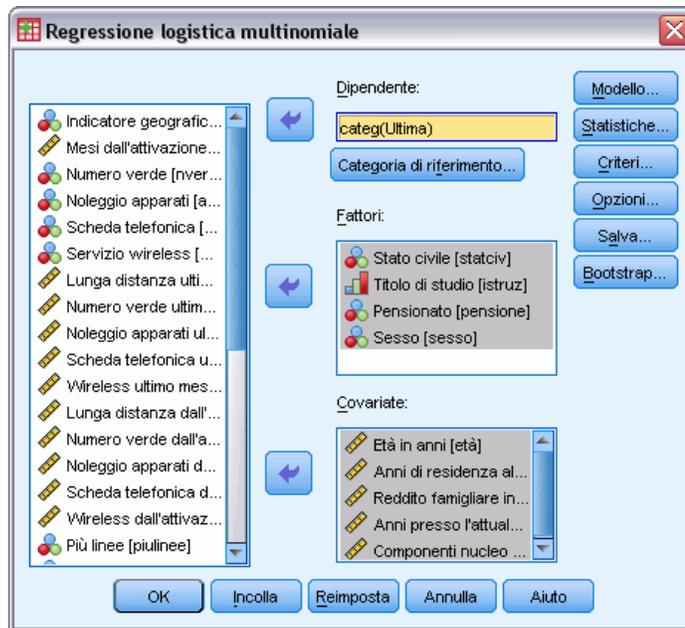
Dati. La variabile dipendente deve essere categoriale. Le variabili indipendenti possono essere fattori o covariate. In genere, i fattori devono essere variabili categoriali e le covariate devono essere variabili continue.

Assunzioni. Si suppone che il rapporto odd di qualsiasi coppia di categorie sia indipendente da tutte le altre categorie. Ad esempio, in base a questa ipotesi, se viene immesso un nuovo prodotto sul mercato, le quote di mercato di tutti gli altri prodotti subiranno una riduzione equamente proporzionale. Inoltre, dato un modello covariata, si suppone che le risposte siano variabili indipendenti multinomiali.

Per ottenere una regressione logistica multinomiale

- Dai menu, scegliere:
Analizza > Regression > Logistica multinomiale...

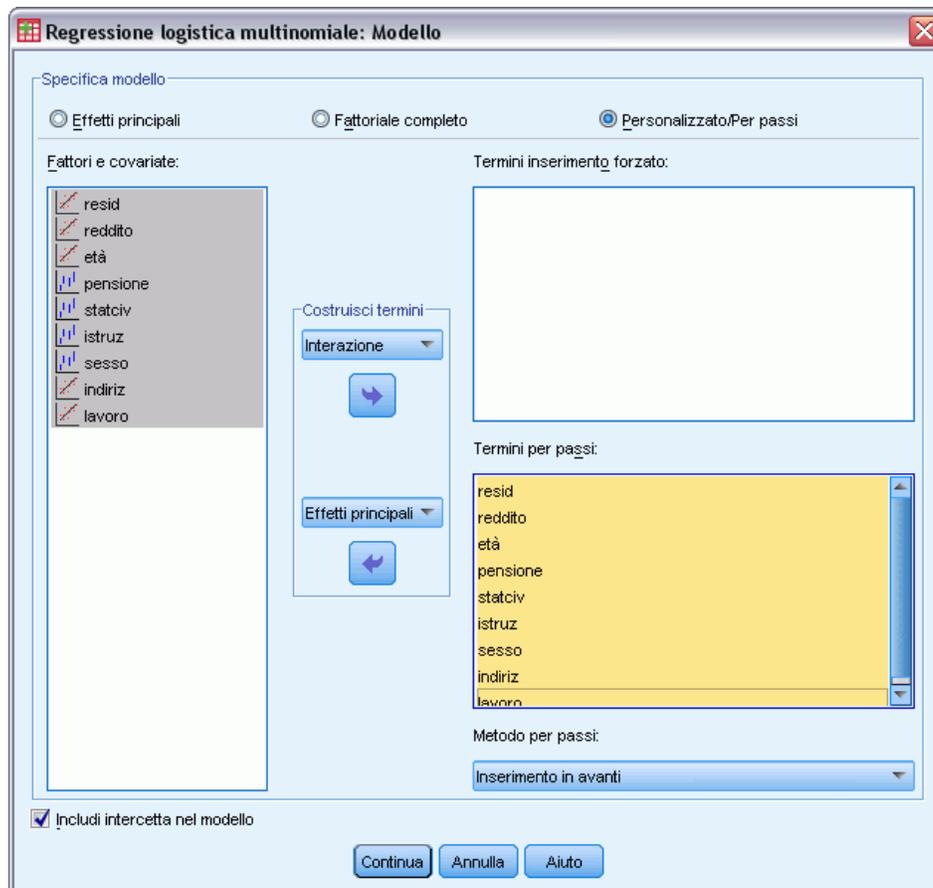
Figura 3-1
Finestra di dialogo *Regressione logistica multinomiale espansa*



- ▶ Selezionare una variabile dipendente.
- ▶ I fattori sono facoltativi e possono essere numerici o categoriali.
- ▶ Le covariate sono facoltative e, se specificate, devono essere valori numerici.

Regressione logistica multinomiale

Figura 3-2
Finestra di dialogo Regressione logistica multinomiale: Modello



Per impostazione predefinita, la procedura Regressione logistica multinomiale crea un modello con fattore ed effetti principali di covariate. Tuttavia in questa finestra di dialogo è possibile specificare un modello personalizzato o una selezione per passi del modello.

Specifica modello. Un modello a effetti principali include gli effetti principali di covariate e fattori, ma non gli effetti di interazione. Un modello fattoriale completo include tutti gli effetti principali dei fattori e tutte le interazioni tra fattori. Non contiene interazioni di covariate. È possibile creare un modello personalizzato per specificare i sottoinsiemi di interazioni di fattori e di interazioni di covariate oppure richiedere la selezione per passi dei termini del modello.

Fattori e covariate. I fattori e le covariate sono elencati.

Termini inserimento forzato. I termini aggiunti all'elenco di inserimento forzato vengono sempre inclusi nel modello.

Termini stepwise. I termini aggiunti all'elenco stepwise vengono inclusi nel modello in base alla selezione di uno dei metodi stepwise seguenti:

- **Inserimento in avanti.** Per l'avvio di questo metodo non viene utilizzato alcun termine stepwise del modello. A ogni passo viene aggiunto il termine più significativo al modello fino ad escludere dal modello i soli termini stepwise il cui contributo al modello non risulta statisticamente significativo.
- **Eliminazione all'indietro.** Per l'avvio di questo modello vengono inseriti tutti i termini specificati nell'elenco stepwise del modello. A ogni passo viene rimosso dal modello il termine stepwise meno significativo fino a mantenere i soli termini stepwise il cui contributo al modello risulta statisticamente significativo.
- **Stepwise in avanti.** Per l'avvio di questo metodo viene utilizzato il modello che sarebbe selezionato dal metodo dell'inserimento in avanti. Dopo l'avvio, l'algoritmo esegue in modo alternato l'eliminazione all'indietro dei termini stepwise del modello e l'inserimento in avanti dei termini esclusi dal modello. Il processo continua fino a quando nessuno termine soddisfa i criteri di inserimento o rimozione.
- **Per passi all'indietro.** Per l'avvio di questo metodo viene utilizzato il modello che sarebbe selezionato dal metodo dell'eliminazione all'indietro. Dopo l'avvio, l'algoritmo esegue in modo alternato l'inserimento in avanti dei termini esclusi dal modello e l'eliminazione all'indietro dei termini stepwise del modello. Il processo continua fino a quando nessuno termine soddisfa i criteri di inserimento o rimozione.

Includi l'intercetta nel modello. Consente di includere o escludere un'intercetta per il modello.

Costruisci termini

Per i fattori e le covariate selezionati:

Interazione. Consente di creare il termine di interazione di livello maggiore rispetto a tutte le variabili selezionate.

Effetti principali. Consente di creare un termine di effetti principali per ciascuna variabile selezionata.

Tutti 2-vie. Consente di creare tutte le possibili interazioni a due vie delle variabili selezionate.

Tutti 3-vie. Consente di creare tutte le possibili interazioni a tre vie delle variabili selezionate.

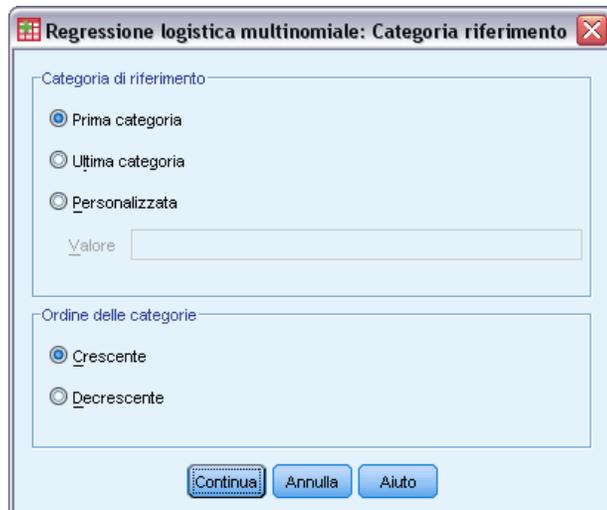
Tutti 4-vie. Consente di creare tutte le possibili interazioni a quattro vie delle variabili selezionate.

Tutti 5-vie. Consente di creare tutte le possibili interazioni a cinque vie delle variabili selezionate.

Regressione logistica multinomiale: Categoria riferimento

Figura 3-3

Finestra di dialogo Regressione logistica multinomiale: Categoria riferimento



Per impostazione predefinita, la procedura Regressione logistica multinomiale definisce l'ultima categoria come categoria di riferimento. Questa finestra di dialogo consente di controllare la categoria di riferimento e la modalità di ordinamento delle categorie.

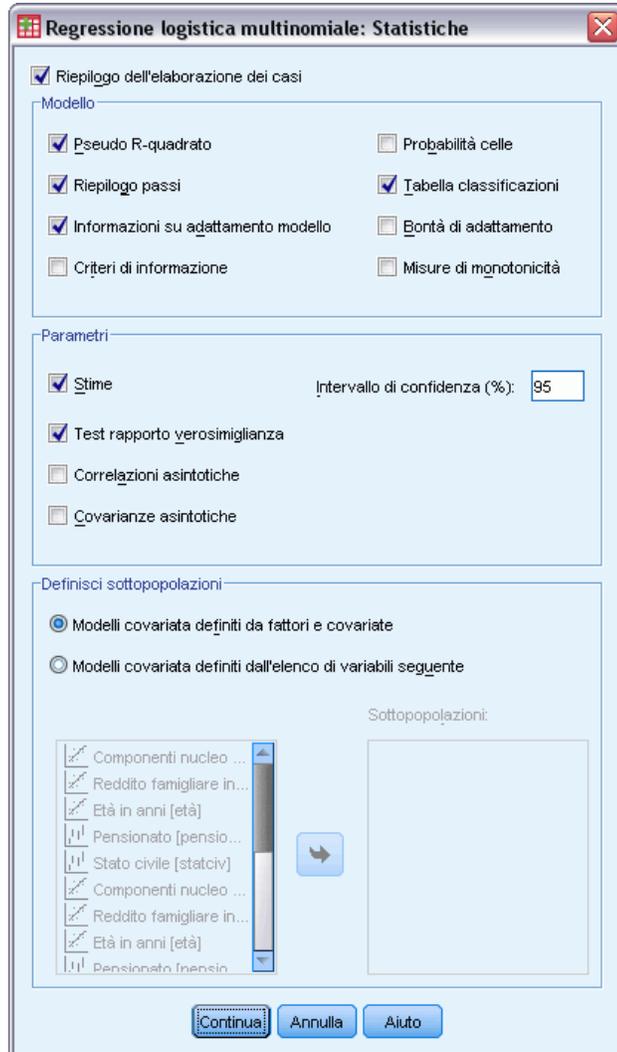
Categoria di riferimento. Specificare la prima o l'ultima categoria oppure una categoria personalizzata.

Ordine delle categorie. Se viene utilizzato l'ordine crescente, il valore più basso definisce la prima categoria e il valore più alto l'ultima. Se viene utilizzato l'ordine decrescente, il valore più alto definisce la prima categoria e il valore più basso l'ultima.

Regressione logistica multinomiale: Statistiche

Figura 3-4

Finestra di dialogo Regressione logistica multinomiale: Statistiche



Per la procedura di regressione logistica multinomiale è possibile specificare le seguenti statistiche:

Riepilogo dei casi. Contiene informazioni sulle variabili categoriali specificate.

Modello. Statistiche del modello complessivo.

- **Pseudo R-quadrato.** Stampa le statistiche R^2 di Cox e Snell, Nagelkerke e McFadden.
- **Riepilogo passi.** Riepiloga gli effetti inseriti o rimossi a ogni passo durante l'esecuzione di un metodo stepwise. Viene generata solo quando viene specificato un modello stepwise nella finestra di dialogo [Modello](#).
- **Inform. su adattam. modello.** Confronta i modelli adattati e quelli con la sola intercetta o nulli.

- **Criteri di informazione.** Questa tabella stampa il Criterio di informazione di Akaike' (AIC) e il Criterio bayesiano di Schwarz' (BIC).
- **Probabilità di cella.** Stampa una tabella delle frequenze osservate e attese (con relativi residui) e le proporzioni per modello covariata e categoria di risposta.
- **Tabella classificazioni.** Stampa una tabella delle risposte osservate confrontate con le risposte attese.
- **Statistiche Chi-quadrato.** Stampa le statistiche chi-quadrato di Pearson e del rapporto di verosimiglianza. Le statistiche vengono elaborate per i modelli covariata determinati usando tutti i fattori e le covariate oppure solo a un sottoinsieme personalizzato di fattori e covariate.
- **Misure di monotonicità.** Visualizza una tabella con informazioni sul numero di coppie concordanti, discordanti e pari merito. In questa tabella sono inoltre visualizzati D di Somers, Gamma di Goodman e Kruskal, tau-a di Kendall e l'indice di concordanza C.

Parametri. Statistiche relative ai parametri del modello.

- **Stime.** Stampa le stime degli effetti del modello, con un livello di confidenza specificato dall'utente.
- **Test del rapporto di verosimiglianza** Stampa i test del rapporto di verosimiglianza per gli effetti del modello parziale. Il test del modello complessivo viene stampato automaticamente.
- **Correlazioni asintotiche.** Stampa la matrice delle correlazioni delle stime dei parametri.
- **Covarianze asintotiche.** Stampa la matrice delle covarianze delle stime dei parametri.

Definisci sottopopolazioni Consente di selezionare un sottoinsieme di fattori e covariate per definire i modelli covariata da utilizzare nel calcolo delle probabilità di cella e dei test della bontà di adattamento.

Regressione logistica multinomiale: Criteri di convergenza

Figura 3-5

Finestra di dialogo Regressione logistica multinomiale: Criteri di convergenza

Regressione logistica multinomiale: Criteri di convergenza

Iterazioni

Massimo numero di iterazioni: 100

Max dimezzamenti: 5

Convergenza log-verosimiglianza: 0

Convergenza parametri: 0.000001

Stampa cronologia iterazioni ogni 1 passi

Controlla separazione dei punti di dati dall'iterazione 20 avanti

Delta: 0 Tolleranza della singolarità: 0.00000001

Continua Annulla Aiuto

Per la regressione logistica multinomiale è possibile specificare i seguenti criteri:

Iterazioni. Consente di specificare il numero massimo di iterazioni per un algoritmo, il numero massimo di passi per i dimezzamenti, le tolleranze di convergenza per modificare la verosimiglianza e i parametri, la frequenza di stampa dell'avanzamento dell'algoritmo iterativo e infine a quale iterazione la procedura inizia la verifica della separazione completa o quasi completa dei dati.

- **Convergenza verosimiglianza.** La convergenza viene presunta se il cambiamento assoluto della funzione di verosimiglianza è minore del valore specificato. Il criterio non è usato se il valore è 0. Specificare un valore non negativo.
- **Convergenza parametri.** La convergenza viene presunta se il cambiamento assoluto delle stime dei parametri è minore del valore specificato. Il criterio non viene utilizzato se il valore specificato è 0.

Delta. Consente di specificare un valore non negativo minore di 1 da aggiungere a tutte le celle vuote delle tavole di contingenza della categoria di risposta per modello covariata. Questa operazione serve a stabilizzare l'algoritmo e ad impedire imperfezioni nelle stime.

Tolleranza della singolarità Consente di specificare la tolleranza utilizzata per controllare le singolarità.

Regressione logistica multinomiale: Criteri di convergenza

Figura 3-6

Finestra di dialogo Regressione logistica multinomiale espansa: Opzioni

Regressione logistica multinomiale: Opzioni

Scala della dispersione

Scala: Nessuna Valore:

Opzioni per passi

Probabilità immissione: Effetti con passi minimi nel modello (per metodi all'indietro)

Test di ingresso: Rapporto di verosimiglianza Effetti con passi minimi nel modello (per metodi in avanti)

Probabilità rimozione:

Test di eliminazione: Rapporto di verosimiglianza

Inserisci e rimuovi termini con vincolo gerarchico

Tratta le covariate come fattori per la determinazione della gerarchia.

Considera solo i termini fattoriali per determinare la gerarchia; è possibile immettere in qualsiasi momento i termini con covariate

All'interno degli effetti covariati, considera solo i termini fattoriali per determinare la gerarchia

Continua Annulla Aiuto

È possibile specificare le seguenti opzioni di Regressione logistica multinomiale:

Scala della dispersione. Consente di specificare il valore della scala di dispersione da utilizzare per correggere la stima della matrice di covarianza parametrica. L'opzione Devianza effettua una stima del valore di scala usando la funzione di devianza (statistica Chi-quadrato del rapporto di verosimiglianza). L'opzione Pearson effettua una stima del valore di scala usando la statistica Chi-quadrato di Pearson. È inoltre possibile specificare un valore di scala personalizzato, che deve essere un valore numerico positivo.

Opzioni per passi. Opzioni che consentono di controllare i criteri statistici quando vengono utilizzati metodi stepwise per la costruzione di un modello. Vengono ignorate a meno che non venga specificato un modello stepwise nella finestra di dialogo [Modello](#).

- **Probab. inserim.** Probabilità del rapporto di verosimiglianza per l'inserimento della variabile. A una probabilità specificata maggiore corrisponde un più facile inserimento della variabile nel modello. Questo criterio viene utilizzato solo quando viene selezionato il metodo di inserimento in avanti, per passi in avanti o per passi all'indietro.
- **Test di immissione.** Metodo utilizzato per immettere termini con i metodi stepwise. Scegliere il test del rapporto di verosimiglianza o il test dei punteggi. Questo criterio viene utilizzato solo quando viene selezionato il metodo di inserimento in avanti, per passi in avanti o per passi all'indietro.
- **Probab. rimozione.** Probabilità del rapporto di verosimiglianza per la rimozione della variabile. A una probabilità specificata maggiore corrisponde una più difficile rimozione della variabile dal modello. Questo criterio viene utilizzato solo quando viene selezionato il metodo di eliminazione all'indietro, per passi in avanti o per passi all'indietro.
- **Test di eliminazione.** Metodo utilizzato per rimuovere termini con i metodi stepwise. Scegliere il test del rapporto di verosimiglianza o il test di Wald. Questo criterio viene utilizzato solo quando viene selezionato il metodo di eliminazione all'indietro, per passi in avanti o per passi all'indietro.
- **Effetti con passi minimi nel modello.** Quando si utilizza il metodo di eliminazione all'indietro o per passi all'indietro, questa opzione consente di specificare il numero minimo di termini da includere nel modello. L'intercetta non viene valutata come termine del modello.
- **Effetti con passi massimi nel modello.** Quando si utilizza il metodo di eliminazione in avanti o per passi in avanti, questa opzione consente di specificare il numero massimo di termini da includere nel modello. L'intercetta non viene valutata come termine del modello.
- **Vincola in modo gerarchico l'inserimento e la rimozione dei termini.** Opzione che consente di scegliere se aggiungere vincoli per l'inclusione dei termini del modello. La gerarchia richiede che vengano innanzitutto inclusi nel modello tutti i termini di ordine inferiore che fanno parte del termine da includere. Ad esempio, se viene utilizzato il requisito di gerarchia, prima di poter aggiungere l'interazione *Stato civile*Sesso* a un modello è necessario che in tale modello siano inclusi i fattori *Stato civile* e *Sesso*. I tre pulsanti di scelta determinano il ruolo delle covariate nella definizione della gerarchia.

Regressione logistica multinomiale: Salva

Figura 3-7

Finestra di dialogo Regressione logistica multinomiale: Salva



La finestra di dialogo Salva consente di salvare le variabili nel file di lavoro e di esportare le informazioni sul modello in un file esterno.

Variabili salvate:

- **Probabilità di risposta stimate.** Probabilità stimate per la classificazione di un modello fattore/covariata nelle categorie di risposta. Ci sono tante probabilità stimate quante sono le categorie della variabile di risposta; ne verranno salvate fino a 25.
- **Categoria prevista.** Categoria di risposta con la maggiore probabilità stimata per un modello fattore o covariata.
- **Probabilità di categoria prevista.** Il massimo delle probabilità di risposta stimate.
- **Probabilità di categoria reale.** Probabilità stimata di classificazione di un modello fattore o covariata nella categoria osservata.

Esporta informazioni modello in file XML. Le stime dei parametri e, se si desidera, le relative covarianze vengono esportati nel file specificato in formato XML (PMML). È possibile utilizzare questo file di modello per applicare le informazioni del modello ad altri file di dati per il calcolo del punteggio.

Opzioni aggiuntive del comando NOMREG

Il linguaggio della sintassi dei comandi consente inoltre di:

- Specificare la categoria di riferimento della variabile dipendente.
- Includere casi con valori utente non validi.
- Personalizzare i test di ipotesi specificando ipotesi nulle come combinazioni lineari dei parametri.

Vedere *Command Syntax Reference* per informazioni dettagliate sulla sintassi.

Analisi Probit

Questa procedura consente di misurare la relazione tra l'intensità dello stimolo e la proporzione di casi che offrono una determinata risposta allo stimolo. È utile per le situazioni in cui si dispone di un output dicotomico che verosimilmente è influenzato o determinato dai livelli di qualche variabile indipendente e ben si adatta ai dati sperimentali. Questa procedura consentirà all'utente di stimare l'intensità di uno stimolo richiesto per indurre una certa proporzione di risposte, quale il dosaggio medio effettivo.

Esempio. Quanto è efficace un nuovo pesticida per l'eliminazione delle formiche e qual'è la giusta concentrazione da utilizzare? È possibile svolgere un esperimento in cui si espongono dei campioni di formiche a diverse concentrazioni di pesticida e quindi registrare il numero di formiche uccise e il numero di formiche esposte. Applicando l'analisi probit a tali dati, è possibile determinare l'intensità della relazione tra la concentrazione e l'eliminazione e quindi determinare qual'è la giusta concentrazione di pesticida per essere sicuri di eliminare, ad esempio, il 95% delle formiche esposte.

Statistiche. I coefficienti di regressione e gli errori standard, intercetta ed errore standard, bontà di adattamento chi-quadrato di Pearson, frequenze osservate e attese e intervalli di confidenza per livelli efficaci di variabili indipendenti. Grafici: grafici a risposta trasformati.

Questa procedura si basa sugli algoritmi proposti e implementati in NPSOL® da Gill, Murray, Saunders e Wright per calcolare le stime dei parametri del modello.

Dati. Per ogni valore della variabile indipendente (o per ogni combinazione di valori per variabili indipendenti multiple), la variabile di risposta deve rappresentare il numero dei casi con i valori che dimostrano la risposta di interesse e il totale osservato deve essere il numero globale di casi con tali valori per la variabile indipendente. La variabile fattore deve essere categoriale, codificata come intera.

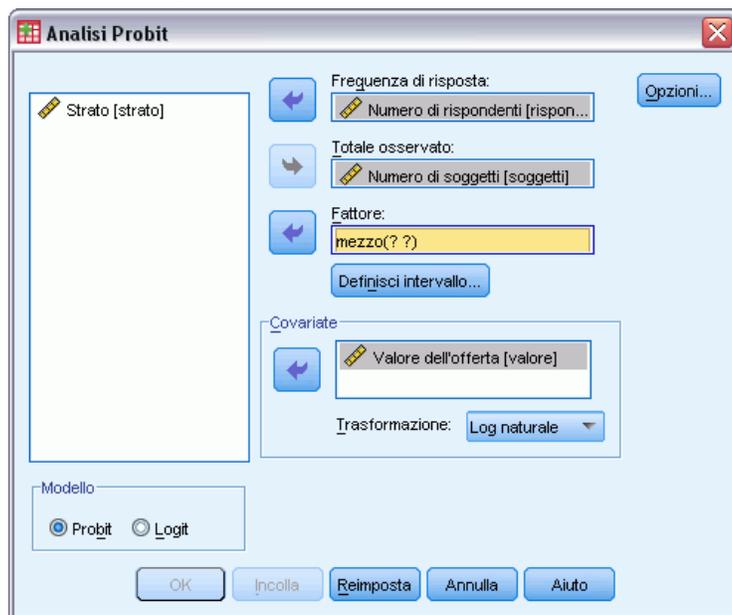
Assunzioni. Le osservazioni devono essere indipendenti. Se si dispone di un ampio numero di valori per le variabili indipendenti relative al numero di osservazioni, come nel caso di studi basati sull'osservazione, è possibile che le statistiche chi-quadrato e bontà di adattamento non siano valide.

Procedure correlate. L'analisi probit è strettamente correlata alla regressione logistica; infatti, se si sceglie la trasformazione logit, questa procedura calcolerà essenzialmente una regressione logistica. In generale, l'analisi probit risulta appropriata per gli esperimenti progettati, mentre la regressione logistica è più appropriata per gli studi basati sull'osservazione. Le differenze nell'output riflettono la diversa enfasi. La procedura di analisi probit riporta le stime dei valori effettivi per i tassi di risposta (incluso il dosaggio effettivo medio), mentre la procedura di regressione logistica riporta le stime dei rapporti odd per le variabili indipendenti.

Per ottenere un'analisi Probit

- Dai menu, scegliere:
Analizza > Regressione > Probit...

Figura 4-1
Finestra di dialogo Analisi Probit



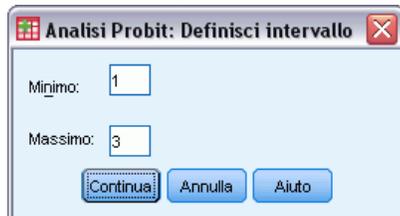
- Selezionare una frequenza di risposta. Questa variabile indica il numero di casi che dimostrano una risposta al test dello stimolo. I valori di questa variabile non possono essere negativi.
- Selezionare un totale osservato. Questa variabile indica il numero di casi a cui è stato applicato lo stimolo. I valori di questa variabile non possono essere negativi e non possono essere minori dei valori delle frequenze di risposta per ogni caso.

In alternativa, è possibile selezionare una variabile fattore. Se lo si desidera, fare clic su Definisci intervallo per definire i gruppi.

- Selezionare una o più covariate. Questa variabile contiene il livello dello stimolo applicato ad ogni osservazione. Se si desidera trasformare la covariata, selezionare una trasformazione dall'elenco a discesa Trasformazione. Se non viene applicata alcuna trasformazione ed esiste un gruppo di controllo, quest'ultimo viene incluso nell'analisi.
- Selezionare il modello Probit o Logit.
 - **Modello Probit.** Applica la trasformazione Probit (l'inversa della funzione di distribuzione cumulata normale standard) ai tassi di risposta.
 - **Modello logit.** Applica la trasformazione logit (rapporti logaritmici) ai tassi di risposta.

Analisi probit: Definisci intervallo

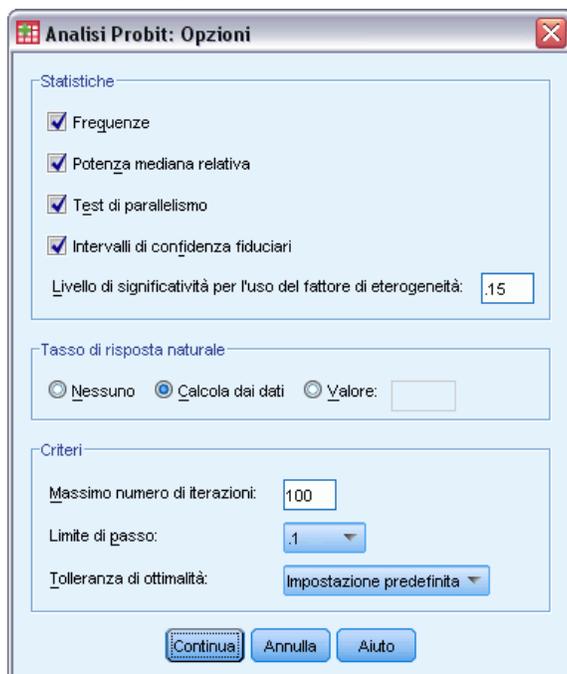
Figura 4-2
Finestra di dialogo Analisi Probit: Definisci intervallo



Consente di specificare i livelli della variabile fattore che verrà analizzata. I livelli devono essere codificati come interi consecutivi e tutti i livelli specificati nell'intervallo verranno analizzati.

Analisi probit: Opzioni

Figura 4-3
Finestra di dialogo Analisi probit: Opzioni



È possibile specificare le opzioni per l'analisi probit:

Statistiche. Consente di richiedere le seguenti statistiche riassuntive: Frequenze, Potenza mediana relativa, Test di parallelismo e Intervalli di confidenza fiduciari.

- **Potenza mediana relativa.** Visualizza il rapporto fra le potenze mediane di ogni coppia di livelli del fattore e i relativi intervalli di confidenza al 95% e i relativi intervalli di confidenza al 95%. Non disponibile se non è stato definito un fattore o se il modello contiene più di una covariata.

- **Test di parallelismo.** Un test dell'ipotesi che tutti i livelli di un fattore abbiano una inclinazione comune, ovvero che siano tutti positivamente o tutti negativamente correlati con la variabile dipendente.
- **Intervalli di confidenza fiduciari.** Intervalli di confidenza per il dosaggio richiesto per ottenere una certa probabilità di risposta.

Gli intervalli di confidenza fiduciari e la potenza mediana relativa non sono disponibili se sono state selezionate più covariate. La potenza mediana relativa e il test di parallelismo sono disponibili solo se è stata selezionata una variabile fattore.

Tasso di risposta naturale. Consente di indicare un tasso di risposta naturale anche in assenza dello stimolo. Le alternative disponibili sono Assente, Calcola dai dati o Valore.

- **Calcola dai dati.** Stima il tasso di risposta naturale dai dati di esempio. I dati devono contenere un caso in cui le covariate assumono valore 0 (il livello di controllo). La procedura stima il tasso di risposta naturale usando la proporzione delle risposte per il livello di controllo come valore iniziale.
- **Valore.** Imposta il tasso di risposta naturale del modello (selezionare questa opzione quando il tasso di risposta naturale è noto in anticipo). Il valore da inserire è la proporzione di risposte, ovvero un valore fra 0 e 1 (0,1=10%, 0,2=20% e così via).

Criteri. Consente di controllare i parametri dell'algoritmo di stima del parametro iterativo. È possibile modificare i valori predefiniti per il numero massimo di iterazioni, il limite di passo e la tolleranza di ottimalità.

Opzioni aggiuntive del comando PROBIT

Il linguaggio della sintassi dei comandi consente inoltre di:

- Richiedere un'analisi sui modelli probit e logit.
- Controllare la gestione dei valori mancanti.
- Trasformare le covariate per basi diverse da 10 o logaritmo naturale.

Per informazioni dettagliate sulla sintassi, vedere *Command Syntax Reference*.

Regressione non lineare

La regressione non lineare è un metodo di ricerca di un modello non lineare per le relazioni tra la variabile dipendente e un insieme di variabili indipendenti. A differenza della regressione lineare, che si limita alla stima di modelli lineari, la regressione non lineare è in grado di stimare modelli con relazioni arbitrarie tra variabili indipendenti e dipendenti. Questa operazione viene eseguita utilizzando algoritmi di stima iterativi. Notare che questa procedura non è necessaria per modelli polinomiali semplici del tipo $Y = A + BX^2$. Se si definisce $W = X^2$, si ottiene un modello lineare semplice, $Y = A + BW$, che può essere stimato utilizzando metodi tradizionali quali la procedura Regressione lineare.

Esempio. È possibile prevedere la popolazione in base al tempo? Un grafico a dispersione mostra che sembra esserci una forte relazione tra la popolazione e il tempo, ma la relazione non è lineare e richiede quindi metodi di stima speciali di regressione non lineare. Impostando un'equazione appropriata, quale il modello di crescita logistico della popolazione, è possibile ottenere una stima valida nel modello, che consente di creare previsioni relative alla popolazione per il futuro.

Statistiche. Per ogni iterazione: stime dei parametri e somma dei residui quadrati. Per ogni modello: somma dei quadrati per la regressione, residuo, totale corretto o non corretto, stime dei parametri, errori standard asintotici e matrice di covarianza asintotica delle stime dei parametri.

Nota: La regressione non lineare vincolata si basa su algoritmi proposti e implementati in NPSOL[®] da Gill, Murray, Saunders e Wright per ottenere stime dei parametri del modello.

Dati. Le variabili dipendenti ed indipendenti devono essere quantitative. È necessario che le variabili categoriali, come la religione, l'età o la regione di residenza, siano ricodificate come variabili binarie (fittizie) o altri tipi di variabili di contrasto.

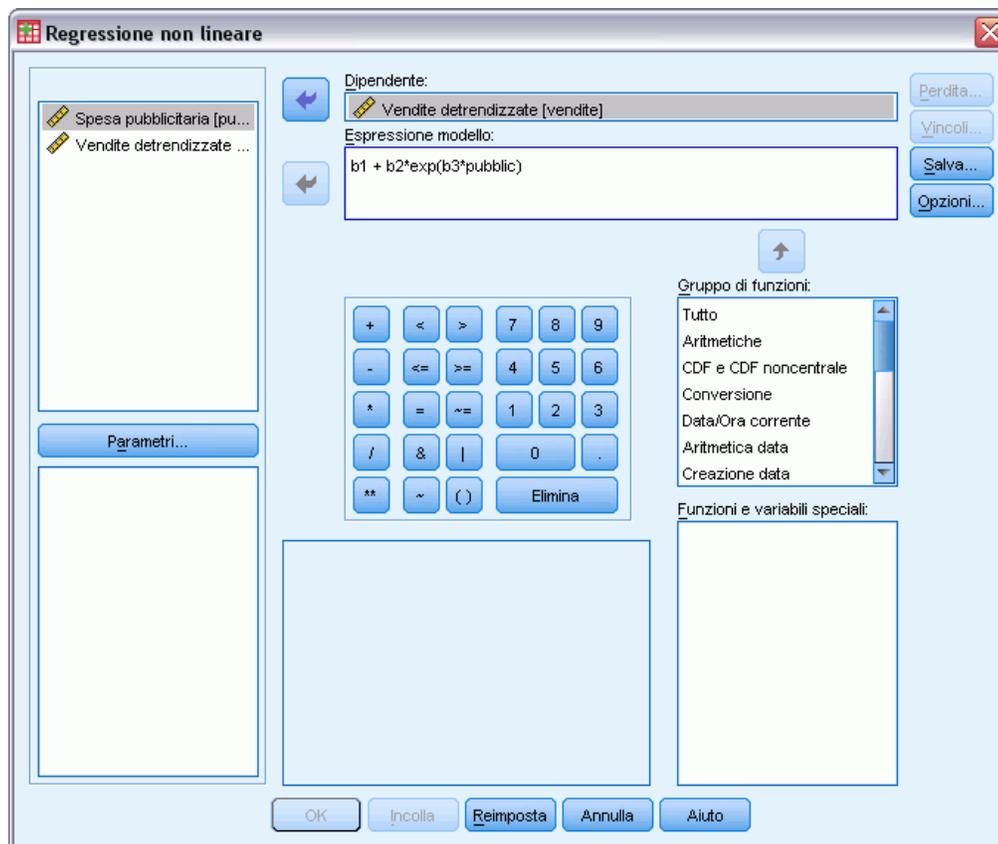
Assunzioni. I risultati sono validi solo se si è specificata una funzione che descrive accuratamente la relazione tra le variabili dipendenti e indipendenti. È inoltre molto importante scegliere buoni valori iniziali. Anche se è stata specificata la forma funzionale corretta del modello, se si utilizzano valori iniziali non appropriati è possibile che il modello non esegua la convergenza oppure che si ottenga una soluzione ottimale a livello locale piuttosto che una soluzione ottimale a livello globale.

Procedure correlate. Molti modelli che sembrano non lineari possono essere trasformati in modelli lineari, che possono essere analizzati utilizzando la procedura di regressione lineare. Se non si è sicuri del modello appropriato da utilizzare, la procedura Stima di curve può contribuire a identificare le relazioni funzionali utili nei dati disponibili.

Per ottenere un'analisi Regressione non lineare

- Dai menu, scegliere:
Analizza > Regressione > Non lineare...

Figura 5-1
Finestra di dialogo *Regressione non lineare*



- ▶ Selezionare una variabile dipendente numerica dall'elenco di variabili nel file dati attivo.
- ▶ Per costruire un'espressione del modello, inserire l'espressione nel campo Espressione modello oppure incollare le componenti (variabili, parametri, funzioni) nel campo.
- ▶ Identificare i parametri nel proprio modello facendo clic su Parametri.

Un modello segmentato (ovvero che assume forme differenti in parti diverse del proprio dominio) deve essere specificato utilizzando la logica condizionale all'interno della singola istruzione di modello.

Logica condizionale (Regressione non lineare)

È possibile specificare un modello segmentato utilizzando la logica condizionale. Per utilizzare la logica condizionale all'interno di un'espressione di modello o di una funzione di perdita, è possibile formare la somma di una serie di termini, uno per ogni condizione. Ogni termine è formato da un'espressione logica (tra parentesi) moltiplicata per l'espressione che dovrebbe risultare quando l'espressione logica è vera.

Ad esempio, si consideri un modello segmentato che è uguale a 0 per $X \leq 0$, X per $0 < X < 1$ e 1 per $X \geq 1$. L'espressione per questo caso è la seguente:

$$(X \leq 0) * 0 + (X > 0 \ \& \ X < 1) * X + (X \geq 1) * 1.$$

Le espressioni logiche tra parentesi definiscono 1 (vero) o 0 (falso). Quindi:

If $X \leq 0$, l'espressione si riduce a $1 * 0 + 0 * X + 0 * 1 = 0$.

Se $0 < X < 1$, si riduce a $0 * 0 + 1 * X + 0 * 1 = X$.

Se $X \geq 1$, si riduce a $0 * 0 + 0 * X + 1 * 1 = 1$.

È possibile creare esempi più complicati tramite la sostituzione di diverse espressioni logiche ed espressioni di risultato. Tenere presente che le ineguaglianze doppie, ad esempio $0 < X < 1$, devono essere scritte come espressioni composte, ad esempio $(X > 0 \ \& \ X < 1)$.

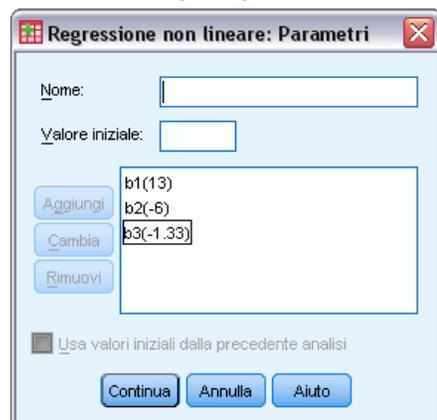
Le variabili stringa possono essere utilizzate all'interno di espressioni logiche:

$$(\text{città} = \text{'New York'}) * \text{costvit} + (\text{città} = \text{'Des Moines'}) * 0.59 * \text{costvit}$$

In questo modo viene fornita un'espressione (il valore della variabile *costvit*) per i cittadini di New York e un'altra (59% di tale valore) per i residenti di Des Moines. Le costanti stringa devono essere racchiuse tra virgolette o apostrofi, come illustrato

Regressione non lineare: Parametri

Figura 5-2
Finestra di dialogo Regressione non lineare: Parametri



I parametri sono parti del modello che vengono stimati dalla procedura di regressione non lineare. I parametri possono essere costanti aggiuntive, coefficienti di moltiplicazione, esponenti o valori utilizzati nelle funzioni di valutazione. Tutti i parametri definiti verranno visualizzati (con i relativi valori iniziali) nell'elenco dei parametri nella finestra di dialogo principale.

Nome. È necessario specificare un nome per ogni parametro. Tale nome deve essere un nome di variabile valido e deve essere il nome utilizzato nell'espressione del modello nella finestra di dialogo principale.

Valore iniziale. Consente di specificare un valore iniziale per il parametro, preferibilmente il più vicino possibile alla soluzione finale attesa. Valori iniziali inappropriati possono causare errori per la convergenza o una convergenza su una soluzione locale (piuttosto che globale) oppure possono essere fisicamente impossibili.

Usa valori iniziali dalla precedente analisi. Se è già stata eseguita una regressione non lineare da questa finestra di dialogo, è possibile selezionare questa opzione per ottenere i valori iniziali dei parametri dai relativi valori nella precedente esecuzione. Ciò consente di continuare la ricerca quando l'algoritmo procede lentamente verso la convergenza. I valori iniziali di partenza verranno ancora visualizzati nell'elenco dei parametri nella finestra di dialogo principale.

Nota: La selezione rimane valida in questa finestra di dialogo per il resto della sessione. Se si modifica il modello, assicurarsi di deseleggerla.

Modelli comuni della Regressione non lineare

La seguente tabella fornisce la sintassi di esempio per molti modelli di regressione non lineare pubblicati. Un modello selezionato casualmente non è in grado di adattarsi ai dati nel modo corretto. È necessario utilizzare i valori iniziali appropriati per i parametri e per alcuni modelli sono necessari vincoli per la convergenza.

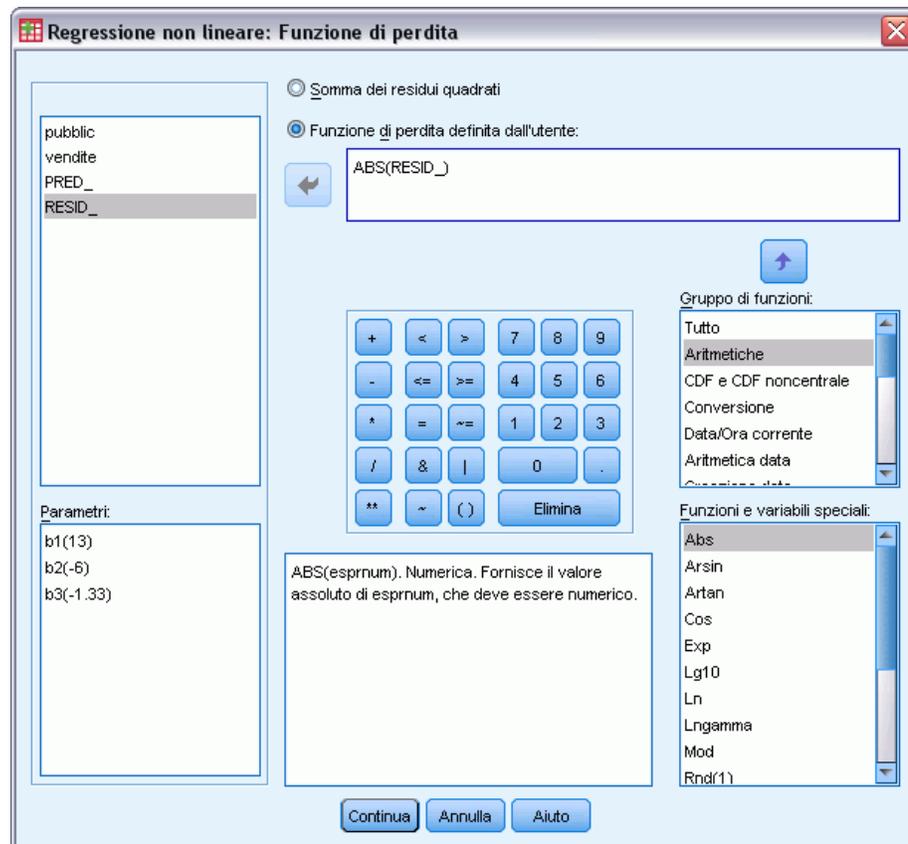
Tabella 5-1
Sintassi di esempio del modello

Nome	Espressione del modello
Regressione asintotica	$b1 + b2 * \exp(b3 * x)$
Regressione asintotica	$b1 - (b2 * (b3 ** x))$
Densità	$(b1 + b2 * x) ** (-1 / b3)$
Gauss	$b1 * (1 - b3 * \exp(-b2 * x ** 2))$
Gompertz	$b1 * \exp(-b2 * \exp(-b3 * x))$
Johnson-Schumacher	$b1 * \exp(-b2 / (x + b3))$
Log modificato	$(b1 + b3 * x) ** b2$
Log logistico	$b1 - \ln(1 + b2 * \exp(-b3 * x))$
Legge dei rendimenti decrescenti di Metcherlich	$b1 + b2 * \exp(-b3 * x)$
Michaelis Menten	$b1 * x / (x + b2)$
Morgan-Mercer-Florin	$(b1 * b2 + b3 * x ** b4) / (b2 + x ** b4)$
Peal-Reed	$b1 / (1 + b2 * \exp(-(b3 * x + b4 * x ** 2 + b5 * x ** 3)))$
Rapporto dei cubi	$(b1 + b2 * x + b3 * x ** 2 + b4 * x ** 3) / (b5 * x ** 3)$
Rapporto dei quadrati	$(b1 + b2 * x + b3 * x ** 2) / (b4 * x ** 2)$
Richards	$b1 / ((1 + b3 * \exp(-b2 * x)) ** (1 / b4))$
Verhulst	$b1 / (1 + b3 * \exp(-b2 * x))$
Von Bertalanffy	$(b1 ** (1 - b4) - b2 * \exp(-b3 * x)) ** (1 / (1 - b4))$
Weibull	$b1 - b2 * \exp(-b3 * x ** b4)$
Riproduzione della densità	$(b1 + b2 * x + b3 * x ** 2) ** (-1)$

Regressione non lineare: Funzione di perdita

Figura 5-3

Finestra di dialogo Regressione non lineare: Funzione di perdita



La **funzione di perdita** nella regressione non lineare è la funzione che viene minimizzata dall'algoritmo. Selezionare Somma dei residui quadrati per minimizzare la somma dei residui quadrati oppure Funzione di perdita definita dall'utente per minimizzare una diversa funzione.

Se si seleziona Funzione di perdita definita dall'utente, è necessario definire la funzione di perdita la cui somma (tra tutti i casi) dovrebbe essere minimizzata dalla scelta dei valori dei parametri.

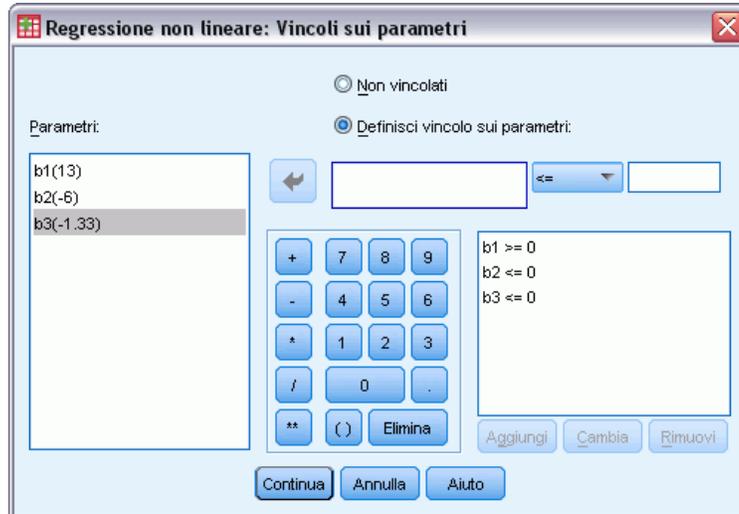
- La maggior parte delle funzioni di perdita interessa la variabile speciale *RESID_* che rappresenta il residuo. La funzione di perdita predefinita, Somma dei residui quadrati, potrebbe essere inserita esplicitamente con *RESID_**2*. Il valore atteso in una funzione di perdita è uguale alla differenza tra la variabile dipendente e il valore residuo.
- È possibile specificare una funzione di perdita condizionale utilizzando la logica condizionale.

È possibile digitare un'espressione nel campo della funzione di perdita definita dall'utente oppure incollare le componenti dell'espressione nel campo. Le costanti stringa devono essere racchiuse tra virgolette o apostrofi e le costanti numeriche devono essere immesse nel formato americano, con il punto come separatore decimale.

Regressione non lineare: Vincoli sui parametri

Figura 5-4

Finestra di dialogo Regressione non lineare: Vincoli sui parametri



Un **vincolo** è una restrizione dei valori disponibili per un parametro durante la ricerca iterativa di una soluzione. Le espressioni lineari vengono valutate prima di eseguire il primo passo, è quindi possibile utilizzare vincoli lineari per impedire i passi che possono causare overflow. Le espressioni non lineari vengono valutate in base al passo.

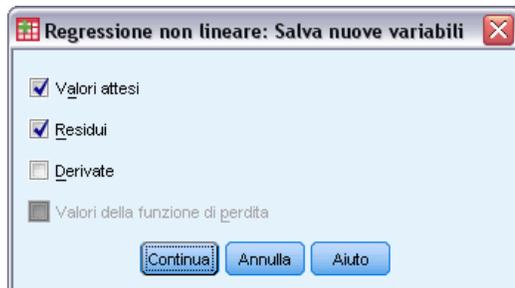
Ogni equazione o ineguaglianza richiede i seguenti elementi:

- Un'espressione che comprende almeno un parametro nel modello. Digitare l'espressione oppure utilizzare la calcolatrice, che consente di incollare i numeri, gli operatori o le parentesi nell'espressione. È possibile digitare i parametri richiesti con il resto dell'espressione oppure incollarli dall'elenco a sinistra. Non è possibile utilizzare variabili ordinarie in un vincolo.
- Uno dei tre operatori logici \leq , $=$ o \geq .
- Una costante numerica, con cui viene confrontata l'espressione utilizzando l'operatore logico. Specificare la costante. Le costanti numeriche devono essere immesse nel formato americano, con il punto come separatore decimale.

Regressione non lineare: Salva nuove variabili

Figura 5-5

Finestra di dialogo Regressione non lineare: Salva nuove variabili



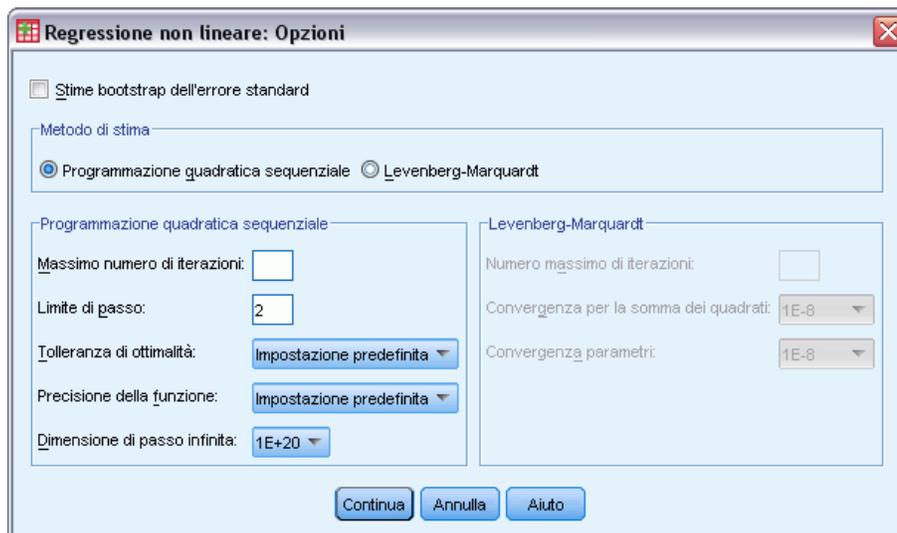
È possibile salvare nuove variabili nel file di dati attivo. Le opzioni disponibili sono Residui, Valori attesi, Derivate e Valori della funzione di perdita. È possibile utilizzare queste variabili in analisi successive per verificare l'adattabilità del modello o per identificare i casi problematici.

- **Residui.** Salva i residui con il nome di variabile resid.
- **Valori attesi.** Salva i valori stimati dal modello. Il nome della variabile è composto dal prefisso pred_ e da un numero progressivo.
- **Derivate.** Viene salvata una derivata per ogni parametro del modello. Il nome delle variabili è composto dal prefisso d. e dai primi sei caratteri del nome del parametro.
- **Valori della funzione di perdita.** Questa opzione è disponibile se è stata specificata una funzione di perdita personalizzata. Il nome della variabile è composto dal prefisso loss_ e da un numero progressivo.

Regressione non lineare: Opzioni

Figura 5-6

Finestra di dialogo Regressione non lineare: Opzioni



Le opzioni consentono di controllare i vari aspetti dell'analisi di regressione non lineare:

Stime bootstrap. Un metodo per stimare l'errore standard di una statistica usando campioni ripetuti dall'insieme di dati originale. Questo risultato viene ottenuto campionando (con sostituzione) per ottenere molti campioni della stessa dimensione dell'insieme di dati originale. L'equazione non lineare è stimata per ciascuno di questi campioni. L'errore standard di ogni stima di parametro viene quindi calcolato come la deviazione standard delle stime di bootstrap. I valori di parametro dei dati originali vengono usati come valori di partenza per ogni campione di bootstrap. È richiesto un algoritmo di programmazione quadratica sequenziale.

Metodo di stima. Consente di selezionare un modello di stima, se possibile. Alcune scelte in questa o in altre finestre di dialogo richiedono l'algoritmo di programmazione quadratica sequenziale. Le alternative disponibili comprendono Programmazione quadratica sequenziale e Levenberg-Marquardt.

- **Programmazione quadratica sequenziale.** Metodo disponibile per modelli vincolati e non vincolati. Viene proposto come metodo predefinito per i modelli vincolati, in presenza di una funzione di perdita personalizzata o se si scelgono le stime bootstrap. È possibile immettere nuovi valori per Massimo numero di iterazioni e Limite di passo ed è possibile modificare la selezione nell'elenco a discesa Tolleranza di ottimalità, Precisione della funzione e Dimensione di passo infinita.
- **Levenberg-Marquardt.** L'algoritmo iterativo predefinito per i modelli non vincolati. Questo metodo non è disponibile per i modelli se è stato specificato un modello vincolato, una funzione di perdita personalizzata o una stima bootstrap. È possibile immettere nuovi valori per Massimo numero di iterazioni ed è possibile modificare la selezione negli elenchi a discesa Convergenza per la somma dei quadrati e Convergenza dei parametri.

Interpretazione dei risultati della regressione non lineare

I problemi di regressione non lineare spesso implicano difficoltà di calcolo:

- La scelta dei valori iniziali per la convergenza di influenza dei parametri. Tentare di scegliere valori iniziali ragionevoli e, se possibile, vicini alla soluzione finale attesa.
- A volte un algoritmo offre prestazioni migliori di un altro in relazione a un particolare problema. Nella finestra di dialogo Opzioni selezionare l'altro algoritmo, se disponibile. Se vengono specificati una funzione di perdita o alcuni tipi di vincoli, non è possibile utilizzare l'algoritmo di Levenberg-Marquardt.
- Quando l'iterazione viene interrotta solo a causa del verificarsi del massimo numero di iterazioni, il modello "finale" probabilmente non rappresenta una buona soluzione. Selezionare Usa i valori iniziali dalla precedente analisi nella finestra di dialogo Parametri per continuare l'iterazione o, ancora meglio, scegliere valori iniziali diversi.
- I modelli che richiedono l'elevazione a potenza di o per dati di grandi dimensioni possono causare overflow o underflow (ovvero numeri troppo grandi o troppo piccoli per essere rappresentati dal computer). A volte è possibile evitare questo problema tramite la scelta di valori iniziali appropriati o tramite l'impostazione di vincoli per i parametri.

Opzioni aggiuntive del comando NLR

Il linguaggio della sintassi dei comandi consente inoltre di:

- Denominare un file da cui leggere i valori iniziali per le stime dei parametri.
- Specificare più di un modello di istruzione e di una funzione di perdita. Ciò facilita la specifica di un modello segmentato.
- Indicare le proprie derivate piuttosto che utilizzare quelle calcolate dal programma.
- Specificare il numero di campioni bootstrap da generare.
- Specificare i criteri aggiuntivi di iterazione, inclusa l'impostazione del valore critico per il controllo delle derivate e la definizione di un criterio di convergenza per la correlazione tra i residui e le derivate.

I criteri aggiuntivi per il comando CNLR (regressione non lineare vincolata) consentono di:

- Specificare il numero massimo delle iterazioni secondarie consentite all'interno di ogni iterazione principale.
- Impostare il valore critico per la verifica delle derivate.
- Impostare il limite di passo.
- Specificare la tolleranza limite per determinare se i valori iniziali sono compresi nei limiti specificati.

Per informazioni dettagliate sulla sintassi, vedere *Command Syntax Reference*.

Stima del peso

Nei modelli di regressione lineare standard si presume che la varianza sia costante all'interno della popolazione in esame. Quando ciò non si verifica, ad esempio quando i casi che hanno valori alti per qualche attributo mostrano una maggiore variabilità dei casi che hanno valori bassi per lo stesso attributo, la regressione lineare che utilizza minimi quadrati ordinari (OLS) non fornisce più stime ottimali del modello. Se è possibile stimare le differenze di variabilità in base a un'altra variabile, la procedura Minimi quadrati ponderati (WLS) può calcolare i coefficienti di un modello di regressione lineare utilizzando i minimi quadrati ponderati (WLS), in modo tale che alle osservazioni più precise (ovvero quelle con una minore variabilità) vengono dati pesi maggiori nella determinazione dei coefficienti di regressione. La procedura Minimi quadrati ponderati (WLS) verifica un intervallo di trasformazioni del peso e indica quale di questi si adatterà al meglio ai dati.

Esempio. Quali sono gli effetti dell'inflazione e della disoccupazione sulle variazioni di prezzo delle scorte? Poiché le merci con valori azionari più alti mostrano spesso una maggiore variabilità di quelle con valori azionari più bassi, i minimi quadrati ordinari non genereranno stime ottimali. La procedura Minimi quadrati ponderati consente di considerare l'effetto del prezzo delle azioni sulla variabilità delle modifiche di prezzo nel calcolo del modello lineare.

Statistiche. Valori di verosimiglianza per ogni potenza del peso della variabile di origine verificata, R multipli, R -quadrati, R -quadrati corretti, tabella ANOVA per il modello WLS, stime di parametri non standardizzate e standardizzate e verosimiglianza per il modello WLS.

Dati. Le variabili dipendenti ed indipendenti devono essere quantitative. È necessario che le variabili categoriali, come la religione, l'età o la regione di residenza, siano ricodificate come variabili binarie (fittizie) o altri tipi di variabili di contrasto. La variabile di ponderazione deve essere quantitativa e deve essere correlata alla variabilità nella variabile dipendente.

Assunzioni. Per ciascun valore della variabile indipendente, la distribuzione della variabile dipendente deve essere normale. La relazione tra la variabile dipendente e ogni variabile indipendente deve essere lineare e tutte le osservazioni devono essere indipendenti. La varianza della variabile dipendente può variare tra i livelli delle variabili indipendenti, ma è necessario che le differenze siano prevedibili in base alla variabile di ponderazione.

Procedure correlate. È possibile utilizzare la procedura Esplora per visualizzare i dati. Questa procedura fornisce i test di normalità e di omogeneità della varianza, nonché visualizzazioni grafiche. Se la variabile dipendente sembra avere una varianza uguale tra i livelli di variabili indipendenti, è possibile utilizzare la procedura di regressione lineare. Se i dati sembrano violare un'ipotesi (quale la normalità), è possibile tentare di trasformarli. Se i dati non sono correlati linearmente e la trasformazione non risulta utile, utilizzare un modello alternativo nella procedura Stima di curve. Se la variabile dipendente è dicotomica, ad esempio se una particolare vendita è completata o se un elemento è difettoso, utilizzare la procedura di regressione logistica. Se la

variabile dipendente è troncata, ad esempio il tempo di sopravvivenza in seguito ad un intervento chirurgico, utilizzare Tavole di sopravvivenza, Kaplan-Meier o Regressione di Cox, disponibili nell'opzione Advanced Statistics. Se i dati non sono indipendenti, ad esempio se si osserva la stessa persona in condizioni diverse, utilizzare la procedura Misure ripetute, disponibile nell'opzione Advanced Statistics.

Per ottenere un'analisi Minimi quadrati ponderati (WLS)

- ▶ Dai menu, scegliere:
Analizza > Regressione > Minimi quadrati ponderati (WLS)...

Figura 6-1
Finestra di dialogo Minimi quadrati ponderati (WLS)

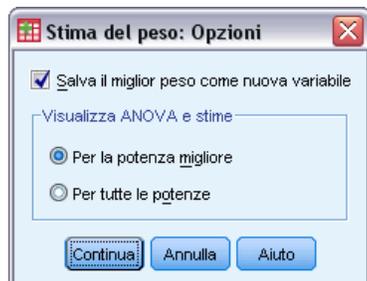


- ▶ Selezionare una variabile dipendente.
- ▶ Selezionare una o più variabili indipendenti.
- ▶ Selezionare la variabile che è l'origine di eteroschedeasticità come variabile di ponderazione.
 - **Variabile di ponderazione.** Le variabili vengono ponderate in base al reciproco di questa variabile elevato a una potenza. L'equazione di regressione è calcolata per ognuna delle potenze incluse nell'intervallo richiesto e indica la potenza che massimizza la funzione di verosimiglianza.
 - **Intervallo di potenza.** Viene utilizzato insieme alla variabile di ponderazione per calcolare i pesi. Verrà calcolata un'equazione di regressione per ciascuna potenza inclusa nell'intervallo specificato. L'intervallo può variare al massimo fra -6,5 e +7,5 compresi. Il passo determina l'incremento utilizzato fra un valore e il successivo. Il numero totale di potenze nell'intervallo è limitato a 150.

Minimi quadrati ponderati (WLS): Opzioni

Figura 6-2

Finestra di dialogo Minimi quadrati ponderati (WLS): Opzioni



È possibile specificare le opzioni per l'analisi Minimi quadrati ponderati:

Salva il miglior peso come nuova variabile. Aggiunge la variabile di ponderazione al file attivo. Il nome di questa variabile è WGT_n , dove n è un numero scelto per assegnare un nome univoco alla variabile.

Visualizza ANOVA e stime. Consente di controllare le modalità di visualizzazione delle statistiche nell'output. Le alternative disponibili sono Per la potenza migliore e Per tutte le potenze.

Opzioni aggiuntive del comando WLS

Il linguaggio della sintassi dei comandi consente inoltre di:

- Fornire un valore singolo per la potenza.
- Specificare un elenco di potenze oppure combinare un intervallo di valori con un elenco di potenze.

Per informazioni dettagliate sulla sintassi, vedere *Command Syntax Reference*.

Minimi quadrati a 2 stadi

Nei modelli di regressione lineare standard si presume che gli errori nella variabile dipendente non siano correlati con le variabili indipendenti. Quando ciò non si verifica, ad esempio quando le relazioni tra variabili sono bidirezionali, la regressione lineare che utilizza i minimi quadrati ordinari (OLS) non fornisce più stime di modello ottimali. La regressione dei minimi quadrati a due stadi utilizza variabili strumentali non correlate ai termini di errore per il calcolo dei valori stimati con stimatori problematici (il primo stadio), quindi utilizza i valori calcolati per effettuare la stima del modello di regressione lineare della variabile dipendente (secondo stadio). Poiché i valori calcolati sono basati su variabili che non sono correlate con gli errori, i risultati del modello a due stadi sono ottimali.

Esempio. La domanda di merci è legata al relativo prezzo e ai redditi dei consumatori? La difficoltà implicita in questo modello è che il prezzo e la domanda hanno effetti reciproci. Ossia, il prezzo può influenzare la domanda e la domanda può influenzare il prezzo. Un modello di regressione dei minimi quadrati a due stadi può utilizzare i redditi dei consumatori e i prezzi ritardati per calcolare un'approssimazione del prezzo non correlata con gli errori di misurazione della domanda. Questa approssimazione viene sostituita al prezzo nel modello specificato originale, che viene quindi stimato.

Statistiche. Per ciascun modello: coefficienti di regressione standardizzati e non standardizzati, R multiplo, R^2 , R^2 corretto, errore standard della stima, tabella di analisi della varianza, valori attesi e residui. Inoltre, gli intervalli di confidenza al 95% per ogni coefficiente di regressione e le matrici di correlazione e covarianza delle stime dei parametri.

Dati. Le variabili dipendenti ed indipendenti devono essere quantitative. È necessario che le variabili categoriali, come la religione, l'età o la regione di residenza, siano ricodificate come variabili binarie (fittizie) o altri tipi di variabili di contrasto. Le variabili esplicative **endogene** devono essere quantitative e non categoriali.

Assunzioni. Per ciascun valore della variabile indipendente, la distribuzione della variabile dipendente deve essere normale. La varianza della distribuzione della variabile dipendente deve essere costante per tutti i valori della variabile indipendente. La relazione tra la variabile dipendente e ogni variabile indipendente deve essere lineare.

Procedure correlate. Se si ritiene che nessuna delle variabili stimatore sia correlata con gli errori nella variabile dipendente, è possibile utilizzare la procedura Regressione lineare. Se invece si ritiene che i dati violino una delle ipotesi, ad esempio la normalità o la varianza di costante, è possibile tentare di trasformarli. Se i dati non sono correlati linearmente e la trasformazione non risulta utile, utilizzare un modello alternativo nella procedura Stima di curve. Se la variabile dipendente è dicotomica, ad esempio una vendita completata o meno, utilizzare la procedura Regressione logistica. Se i dati non sono indipendenti, ad esempio se si osserva la stessa persona

in condizioni diverse, utilizzare la procedura Misure ripetute, disponibile nell'opzione Modelli avanzati.

Per ottenere un'analisi di regressione Minimi quadrati a 2 stadi

- ▶ Dai menu, scegliere:
Analizza > Regressione > Minimi quadrati a 2 stadi...

Figura 7-1

Finestra di dialogo Minimi quadrati a 2 stadi

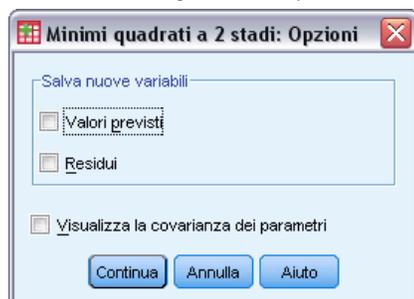


- ▶ Selezionare una variabile dipendente.
- ▶ Selezionare una o più variabili esplicative (stimatore).
- ▶ Selezionare una o più variabili strumentali.
 - **Strumentali.** Le variabili utilizzate per stimare i valori previsti delle variabili endogene nel primo stadio dell'analisi dei minimi quadrati a due stadi. Le stesse variabili possono figurare sia nella casella di riepilogo Esplicative che in quella Strumentali. Il numero di variabili strumentali deve essere almeno uguale a quello di variabili esplicative. Se le variabili strumentali ed esplicative sono esattamente le stesse, il risultato dell'analisi sarà lo stesso di una regressione lineare.

Le variabili esplicative non definite come strumentali sono considerate endogene. In genere tutte le variabili esogene dell'elenco Esplicative sono definite anche come variabili strumentali.

Minimi quadrati a 2 stadi: Opzioni

Figura 7-2
Finestra di dialogo Minimi quadrati a 2 stadi: Opzioni



È possibile selezionare le seguenti opzioni per l'analisi:

Salva nuove variabili. Consente di aggiungere nuove variabili al file attivo. Le opzioni disponibili sono Valori previsti e Residui.

Visualizza la covarianza dei parametri. Consente di stampare la matrice di covarianza delle stime dei parametri.

Opzioni aggiuntive del comando 2SLS

Il linguaggio della sintassi dei comandi consente inoltre di eseguire la stima contemporanea di più equazioni. Per informazioni dettagliate sulla sintassi, vedere *Command Syntax Reference*.

Schemi di codifica delle variabili categoriali

In molte procedure, è possibile richiedere la sostituzione automatica di una variabile categoriale indipendente con un insieme di variabili di contrasto, che verranno poi inserite o rimosse da un'equazione come blocco. È possibile specificare la modalità di codifica delle variabili di contrasto, solitamente nel sottocomando `CONTRAST`. In questa appendice viene spiegata e illustrata la diversità di funzionamento dei vari tipi di contrasto richiesti nel sottocomando `CONTRAST`.

Deviazione

Deviazione dalla media principale. In termini di matrici, questi contrasti assumono la forma di:

media	($1/k$	$1/k$...	$1/k$	$1/k$)
df(1)	($1-1/k$	$-1/k$...	$-1/k$	$-1/k$)
df(2)	($-1/k$	$1-1/k$...	$-1/k$	$-1/k$)
.			.				
.			.				
df(k-1)	($-1/k$	$-1/k$...	$1-1/k$	$-1/k$)

dove k è il numero di categorie per la variabile indipendente e l'ultima categoria viene omessa per impostazione predefinita. Ad esempio, i contrasti di deviazione per una variabile indipendente con tre categorie sono i seguenti:

($1/3$	$1/3$	$1/3$)
($2/3$	$-1/3$	$-1/3$)
($-1/3$	$2/3$	$-1/3$)

Per omettere una categoria diversa dall'ultima, specificare tra parentesi il numero della categoria omessa dopo la parola chiave `DEVIATION`. Ad esempio, il seguente sottocomando consente di recuperare le deviazioni per la prima e la terza categoria, omettendo la seconda:

`/CONTRAST (FACTOR) =DEVIATION (2)`

Si supponga che il *fattore* abbia tre categorie. La matrice di contrasto risultante sarà

$$\begin{pmatrix} 1/3 & 1/3 & 1/3 \\ 2/3 & -1/3 & -1/3 \\ -1/3 & -1/3 & 2/3 \end{pmatrix}$$

Semplice

Contrasti semplici. Confronta ogni livello di un fattore con l'ultimo. La forma generale della matrice è

$$\begin{array}{l} \text{media} \\ \text{df}(1) \\ \text{df}(2) \\ \cdot \\ \cdot \\ \text{df}(k-1) \end{array} \begin{pmatrix} 1/k & 1/k & \dots & 1/k & 1/k \\ 1 & 0 & \dots & 0 & -1 \\ 0 & 1 & \dots & 0 & -1 \\ \cdot & \cdot & & & \\ \cdot & \cdot & & & \\ 0 & 0 & \dots & 1 & -1 \end{pmatrix}$$

dove k è il numero di categorie per la variabile indipendente. Ad esempio, i contrasti semplici per una variabile indipendente con quattro categorie sono i seguenti:

$$\begin{pmatrix} 1/4 & 1/4 & 1/4 & 1/4 \\ 1 & 0 & 0 & -1 \\ 0 & 1 & 0 & -1 \\ 0 & 0 & 1 & -1 \end{pmatrix}$$

Per utilizzare un'altra categoria invece dell'ultima come categoria di riferimento, specificare tra parentesi, dopo la parola chiave `SIMPLE`, il numero di sequenza della categoria di riferimento, che non corrisponderà necessariamente al valore associato alla categoria. Ad esempio, il seguente sottocomando `CONTRAST` ottiene una matrice di contrasto che omette la seconda categoria:

```
/CONTRAST(FACTOR) = SIMPLE(2)
```

Si supponga che il *fattore* abbia quattro categorie. La matrice di contrasto risultante sarà

$$\begin{pmatrix} 1/4 & 1/4 & 1/4 & 1/4 \\ 1 & -1 & 0 & 0 \\ 0 & -1 & 1 & 0 \\ 0 & -1 & 0 & 1 \end{pmatrix}$$

Helmert

Contrasti di Helmert. Confronta le categorie di una variabile indipendente con la media delle categorie successive. La forma generale della matrice è

media	($1/k$	$1/k$...	$1/k$	$1/k$)
df(1)	(1	$-1/(k-1)$...	$-1/(k-1)$	$-1/(k-1)$)
df(2)	(0	1	...	$-1/(k-2)$	$-1/(k-2)$)
.			.				
.			.				
df(k-2)	(0	0	1	$-1/2$	$-1/2$)
df(k-1)	(0	0	...	1	-1)

dove k è il numero di categorie per la variabile indipendente. Ad esempio, una variabile indipendente con quattro categorie avrà una matrice di contrasto di Helmert composta nel seguente modo:

($1/4$	$1/4$	$1/4$	$1/4$)
(1	$-1/3$	$-1/3$	$-1/3$)
(0	1	$-1/2$	$-1/2$)
(0	0	1	-1)

Differenza

Differenza o contrasti inversi di Helmert. Confronta le categorie di una variabile indipendente con la media delle categorie precedenti della variabile. La forma generale della matrice è

media	($1/k$	$1/k$	$1/k$...	$1/k$)
df(1)	(-1	1	0	...	0)
df(2)	($-1/2$	$-1/2$	1	...	0)
.			.				
.			.				
df(k-1)	($-1/(k-1)$	$-1/(k-1)$	$-1/(k-1)$...	1)

dove k è il numero di categorie per la variabile indipendente. Ad esempio, i contrasti differenza per una variabile indipendente con quattro categorie sono i seguenti:

($1/4$	$1/4$	$1/4$	$1/4$)
(-1	1	0	0)
($-1/2$	$-1/2$	1	0)
($-1/3$	$-1/3$	$-1/3$	1)

Polinomiale

Contrasti polinomiali ortogonali. Il primo grado di libertà include l'effetto lineare di tutte le categorie, il secondo grado di libertà l'effetto quadratico, il terzo grado di libertà l'effetto cubico e così via per gli effetti di ordine più alto.

È possibile specificare la distanza tra i diversi livelli del trattamento misurata in base alla variabile categoriale. La distanza uguale, per impostazione predefinita se si omette la metrica, può essere specificata con numeri interi consecutivi da 1 a k , dove k è il numero di categorie. Se la variabile relativa al *medicinale* ha tre categorie, il sottocomando

```
/CONTRAST (DRUG) = POLYNOMIAL
```

è uguale a

```
/CONTRAST (DRUG) = POLYNOMIAL (1, 2, 3)
```

Non è tuttavia sempre necessaria una distanza uguale. Ad esempio, si supponga che *DRUG* rappresenti diversi dosaggi di medicinale somministrati a tre gruppi. Se il dosaggio somministrato al secondo gruppo è pari al doppio del primo gruppo e il dosaggio somministrato al terzo gruppo è il triplo di quello del primo gruppo, le categorie del trattamento presentano la stessa distanza. La metrica appropriata per questa situazione è costituita da numeri interi consecutivi:

```
/CONTRAST (DRUG) = POLYNOMIAL (1, 2, 3)
```

Se, tuttavia, il dosaggio somministrato al secondo gruppo è quattro volte più grande rispetto a quello del primo gruppo e il dosaggio somministrato al terzo gruppo è sette volte quello del primo gruppo, la metrica appropriata sarà

```
/CONTRAST (DRUG) = POLYNOMIAL (1, 4, 7)
```

In ambedue i casi, il risultato della specifica del contrasto è che il primo grado di libertà per la variabile *drug* include l'effetto lineare dei livelli di dosaggio e il secondo grado di libertà include l'effetto quadratico.

I contrasti polinomiali si rivelano particolarmente utili nei test dei trend e nelle indagini sulla natura delle superfici di risposta. È inoltre possibile utilizzare i contrasti polinomiali per eseguire l'adattamento della curva non lineare, ad esempio nella regressione curvilineare.

Ripetuto

Confronta i livelli adiacenti di una variabile indipendente. La forma generale della matrice è

media	(1/k	1/k	1/k	...	1/k	1/k)
df(1)	(1	-1	0	...	0	0)
df(2)	(0	1	-1	...	0	0)
.
.
df(k-1)	(0	0	0	...	1	-1)

dove k è il numero di categorie per la variabile indipendente. Ad esempio, i contrasti ripetuti per una variabile indipendente con quattro categorie sono i seguenti:

$$\begin{pmatrix} 1/4 & 1/4 & 1/4 & 1/4 \\ 1 & -1 & 0 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & -1 \end{pmatrix}$$

Questi contrasti risultano utili per l'analisi del profilo e in qualsiasi caso siano necessari punteggi di differenza.

Contrasto speciale

Un contrasto definito dall'utente. Consente l'immissione di contrasti speciali come matrici quadrate con un numero di righe e di colonne pari al numero delle categorie della variabile indipendente. Per MANOVA e LOGLINEAR, la prima riga inserita corrisponde sempre all'effetto medio, o costante, e rappresenta l'insieme di pesi che determina la media delle altre variabili in base alla variabile specificata. In genere, questo contrasto è un vettore.

Le altre righe della matrice contengono i contrasti speciali che indicano i confronti desiderati tra le categorie della variabile. Solitamente, risultano più utili i contrasti ortogonali. I contrasti ortogonali sono statisticamente indipendenti e non ridondanti. I contrasti sono ortogonali se:

- Per ogni riga, la somma dei coefficienti di contrasto è pari a 0.
- La somma dei prodotti dei coefficienti corrispondenti relativi a tutte le coppie di righe non congiunte corrisponde a 0.

Si supponga, ad esempio, che il trattamento presenti quattro livelli e che si desideri confrontarli tra loro. Un contrasto speciale appropriato è

$(1 \quad 1 \quad 1 \quad 1)$	pesi per il calcolo della media
$(3 \quad -1 \quad -1 \quad -1)$	confronto tra il livello 1 e i livelli da 2 a 4
$(0 \quad 2 \quad -1 \quad -1)$	confronto tra il livello 2 e i livelli da 3 a 4
$(0 \quad 0 \quad 1 \quad -1)$	confronto tra il livello 3 e il livello 4

Il confronto viene specificato tramite il seguente sottocomando CONTRAST per MANOVA, LOGISTIC REGRESSION e COXREG:

```
/CONTRAST (TREATMNT)=SPECIAL ( 1 1 1 1
                                3 -1 -1 -1
                                0 2 -1 -1
                                0 0 1 -1 )
```

Per LOGLINEAR, è necessario specificare:

```
/CONTRAST (TREATMNT)=BASIS SPECIAL ( 1 1 1 1
                                       3 -1 -1 -1
                                       0 2 -1 -1
                                       0 0 1 -1 )
```

La somma di ogni riga ad eccezione della riga relativa alla media è pari a 0. Anche la somma dei prodotti di ogni coppia di righe non congiunte è pari a 0:

$$\text{Righe 2 e 3:} \quad (3)(0) + (-1)(2) + (-1)(-1) + (-1)(-1) = 0$$

$$\text{Righe 2 e 4:} \quad (3)(0) + (-1)(0) + (-1)(1) + (-1)(-1) = 0$$

$$\text{Righe 3 e 4:} \quad (0)(0) + (2)(0) + (-1)(1) + (-1)(-1) = 0$$

Non è necessario che i contrasti speciali siano ortogonali. Tuttavia, non possono essere combinazioni lineari gli uni degli altri. In caso lo siano, la procedura segnala la dipendenza lineare e l'elaborazione viene interrotta. I contrasti di Helmert, di differenza e polinomiali sono contrasti ortogonali.

Indicatore

Codifica della variabile dell'indicatore. Denominata anche codifica fittizia, non è disponibile in LOGLINEAR o MANOVA. Il numero delle nuove variabili codificate è $k-1$. La codifica dei casi nella categoria di riferimento è pari a 0 per tutte le variabili $k-1$. La codifica di un caso nella categoria i è 0 per tutte le variabili dell'indicatore, ad eccezione della variabile i per la quale è 1.

Notices

Licensed Materials – Property of SPSS Inc., an IBM Company. © Copyright SPSS Inc. 1989, 2010.

Patent No. 7,023,453

The following paragraph does not apply to the United Kingdom or any other country where such provisions are inconsistent with local law: SPSS INC., AN IBM COMPANY, PROVIDES THIS PUBLICATION “AS IS” WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Some states do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you.

This information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. SPSS Inc. may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

Any references in this information to non-SPSS and non-IBM Web sites are provided for convenience only and do not in any manner serve as an endorsement of those Web sites. The materials at those Web sites are not part of the materials for this SPSS Inc. product and use of those Web sites is at your own risk.

When you send information to IBM or SPSS, you grant IBM and SPSS a nonexclusive right to use or distribute the information in any way it believes appropriate without incurring any obligation to you.

Information concerning non-SPSS products was obtained from the suppliers of those products, their published announcements or other publicly available sources. SPSS has not tested those products and cannot confirm the accuracy of performance, compatibility or any other claims related to non-SPSS products. Questions on the capabilities of non-SPSS products should be addressed to the suppliers of those products.

This information contains examples of data and reports used in daily business operations. To illustrate them as completely as possible, the examples include the names of individuals, companies, brands, and products. All of these names are fictitious and any similarity to the names and addresses used by an actual business enterprise is entirely coincidental.

COPYRIGHT LICENSE:

This information contains sample application programs in source language, which illustrate programming techniques on various operating platforms. You may copy, modify, and distribute these sample programs in any form without payment to SPSS Inc., for the purposes of developing,

using, marketing or distributing application programs conforming to the application programming interface for the operating platform for which the sample programs are written. These examples have not been thoroughly tested under all conditions. SPSS Inc., therefore, cannot guarantee or imply reliability, serviceability, or function of these programs. The sample programs are provided “AS IS”, without warranty of any kind. SPSS Inc. shall not be liable for any damages arising out of your use of the sample programs.

Trademarks

IBM, the IBM logo, and [ibm.com](http://www.ibm.com) are trademarks of IBM Corporation, registered in many jurisdictions worldwide. A current list of IBM trademarks is available on the Web at <http://www.ibm.com/legal/copytrade.shtml>.

SPSS is a trademark of SPSS Inc., an IBM Company, registered in many jurisdictions worldwide.

Adobe, the Adobe logo, PostScript, and the PostScript logo are either registered trademarks or trademarks of Adobe Systems Incorporated in the United States, and/or other countries.

Intel, Intel logo, Intel Inside, Intel Inside logo, Intel Centrino, Intel Centrino logo, Celeron, Intel Xeon, Intel SpeedStep, Itanium, and Pentium are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries.

Linux is a registered trademark of Linus Torvalds in the United States, other countries, or both.

Microsoft, Windows, Windows NT, and the Windows logo are trademarks of Microsoft Corporation in the United States, other countries, or both.

UNIX is a registered trademark of The Open Group in the United States and other countries.

Java and all Java-based trademarks and logos are trademarks of Sun Microsystems, Inc. in the United States, other countries, or both.

This product uses WinWrap Basic, Copyright 1993-2007, Polar Engineering and Consulting, <http://www.winwrap.com>.

Other product and service names might be trademarks of IBM, SPSS, or other companies.

Adobe product screenshot(s) reprinted with permission from Adobe Systems Incorporated.

Microsoft product screenshot(s) reprinted with permission from Microsoft Corporation.



- Analisi Probit
 - criteri, 23
 - Definisci intervallo, 23
 - esempio, 21
 - funzioni aggiuntive del comando, 24
 - intervalli di confidenza fiduciari, 23
 - iterazioni, 23
 - potenza mediana relativa, 23
 - statistiche, 21, 23
 - tasso di risposta naturale, 23
 - test del parallelismo, 23

- bontà di adattamento
 - in regressione logistica multinomiale, 16

- categoria di riferimento
 - in regressione logistica multinomiale, 15
- celle con zero osservazioni
 - in regressione logistica multinomiale, 17
- Chi-quadrato di Pearson
 - bontà di adattamento, 16
 - per la stima del valore della scala di dispersione, 18
- classificazione
 - in regressione logistica multinomiale, 11
- contrasti
 - in regressione logistica, 6
- covariate
 - in regressione logistica, 6
- covariate categoriche, 6
- covariate stringa
 - in regressione logistica, 6
- criterio di convergenza
 - in regressione logistica multinomiale, 17
- cronologia iterazioni
 - in regressione logistica multinomiale, 17

- delta
 - come correzione per le celle con zero osservazioni, 17
- differenza in beta
 - in regressione logistica, 7
- dimezzamenti
 - in regressione logistica multinomiale, 17
- Distanza di Cook
 - in regressione logistica, 7

- eliminazione all'indietro
 - in regressione logistica, 5

- funzione di devianza
 - per la stima del valore della scala di dispersione, 18

- Intercetta
 - Includi o Escludi, 13
- intervalli di confidenza
 - in regressione logistica multinomiale, 16
- intervalli di confidenza fiduciari
 - in analisi Probit, 23
- iterazioni
 - in analisi Probit, 23
 - in regressione logistica, 9
 - in regressione logistica multinomiale, 17

- legal notices, 46
- Legge dei rendimenti decrescenti di Metcherlich
 - in regressione non lineare, 28

- matrice di correlazione
 - in regressione logistica multinomiale, 16
- matrice di covarianza
 - in regressione logistica multinomiale, 16
- Minimi quadrati a 2 stadi, 37
 - covarianza dei parametri, 39
 - esempio, 37
 - funzioni aggiuntive del comando, 39
 - salvataggio di nuove variabili, 39
 - statistiche, 37
 - variabili strumentali, 37
- modelli fattoriali completi
 - in regressione logistica multinomiale, 13
- modelli non lineari
 - in regressione non lineare, 28
- modelli personalizzati
 - in regressione logistica multinomiale, 13
- modello a effetti principali
 - in regressione logistica multinomiale, 13
- modello del rapporto dei cubi
 - in regressione non lineare, 28
- modello del rapporto dei quadrati
 - in regressione non lineare, 28
- modello di densità
 - in regressione non lineare, 28
- modello di Gauss
 - in regressione non lineare, 28
- modello di Gompertz
 - in regressione non lineare, 28
- modello di Johnson-Schumacher
 - in regressione non lineare, 28
- modello di Michaelis Menten
 - in regressione non lineare, 28
- modello di Morgan-Mercer-Florin
 - in regressione non lineare, 28

- modello di Peal-Reed
 - in regressione non lineare, 28
- modello di Richards
 - in regressione non lineare, 28
- modello di riproduzione della densità
 - in regressione non lineare, 28
- modello di Verhulst
 - in regressione non lineare, 28
- modello di Von Bertalanffy
 - in regressione non lineare, 28
- modello di Weibull
 - in regressione non lineare, 28
- modello log modificato
 - in regressione non lineare, 28

- potenza mediana relativa
 - in analisi Probit, 23

- R quadrato di Cox e Snell
 - in regressione logistica multinomiale, 16
- R-quadrato di McFadden
 - in regressione logistica multinomiale, 16
- R-quadrato di Nagelkerke
 - in regressione logistica multinomiale, 16
- rapporto di verosimiglianza
 - bontà di adattamento, 16
 - per la stima del valore della scala di dispersione, 18
- regressione asintotica
 - in regressione non lineare, 28
- Regressione lineare
 - Minimi quadrati a 2 stadi, 37
 - minimi quadrati ponderati (WLS), 34
- Regressione logistica, 3
 - binaria, 1
 - coefficienti, 3
 - contrasti, 6
 - covariate categoriche, 6
 - covariate stringa, 6
 - Definisci regola di selezione, 5
 - esempio, 3
 - funzioni aggiuntive del comando, 10
 - Imposta valore, 5
 - iterazioni, 9
 - metodi di selezione delle variabili, 5
 - misure dell'influenza, 7
 - opzioni di visualizzazione, 9
 - probabilità per metodi stepwise., 9
 - residui, 7
 - salvataggio di nuove variabili, 7
 - statistiche, 3
 - statistiche e grafici, 9
 - termine costante, 9
 - Test di bontà dell'adattamento di Hosmer-Lemeshow, 9
 - valore di riferimento, 9
 - valori attesi, 7
- regressione logistica binaria, 1
- Regressione logistica multinomiale, 11, 16
 - categoria di riferimento, 15
 - criteri, 17
 - esportazione di informazioni dei modelli, 20
 - funzioni aggiuntive del comando, 20
 - modelli, 13
 - Salva, 20
 - statistiche, 16
- Regressione non lineare, 25
 - algoritmo di Levenberg-Marquardt, 31
 - derivate, 31
 - esempio, 25
 - funzione di perdita, 29
 - funzioni aggiuntive del comando, 33
 - interpretazione dei risultati, 32
 - logica condizionale, 26
 - metodi di stima, 31
 - modelli non lineari comuni, 28
 - modello segmentato, 26
 - parametri, 27
 - programmazione quadratica sequenziale, 31
 - residuo, 31
 - Salva nuove variabili, 31
 - statistiche, 25
 - stime bootstrap, 31
 - valori attesi, 31
 - valori iniziali, 27
 - vincoli sui parametri, 30
- regressione vincolata
 - in regressione non lineare, 30

- selezione in avanti
 - in regressione logistica, 5
- selezione per passi
 - in regressione logistica, 5
 - in regressione logistica multinomiale, 13
- separazione
 - in regressione logistica multinomiale, 17
- singularità
 - in regressione logistica multinomiale, 17
- Stima del peso, 34
 - cronologia iterazioni, 36
 - esempio, 34
 - funzioni aggiuntive del comando, 36
 - Salva il miglior peso come nuova variabile, 36
 - statistiche, 34
 - verosimiglianza, 34
 - visualizzazione di ANOVA e stime, 36
- stime dei parametri
 - in regressione logistica multinomiale, 16

- tabelle di classificazione
 - in regressione logistica multinomiale, 16
- tabelle di probabilità di cella
 - in regressione logistica multinomiale, 16
- termine costante
 - in regressione lineare, 9

Indice

test del parallelismo

in analisi Probit, 23

Test di bontà dell'adattamento di Hosmer-Lemeshow

in regressione logistica, 9

trademarks, 47

valore della scala di dispersione

in regressione logistica multinomiale, 18

valori d'influenza

in regressione logistica, 7

verosimiglianza

in Minimi quadrati ponderati (WLS), 34

in regressione logistica multinomiale, 16

vincoli sui parametri

in regressione non lineare, 30